The Bethe Permanent of a Non-Negative Matrix

Pascal O. Vontobel

***Abstract*—It has recently been observed that the permanent of a non-negative square matrix, *i.e.*, of a square matrix containing only non-negative real entries, can very well be approximated by solving a certain Bethe free energy function minimization problem with the help of the sum-product algorithm. We call the resulting approximation of the permanent the Bethe permanent.**

In this paper we give reasons why this approach to approximating the permanent works well. Namely, we show that the Bethe free energy function is convex and that the sum-product algorithm finds its minimum efficiently. We also discuss the fact that the permanent is lower bounded by the Bethe permanent, and we comment on potential upper bounds on the permanent based on the Bethe permanent. We also present a combinatorial characterization of the Bethe permanent in terms of permanents of so-called lifted versions of the matrix under consideration.

Moreover, we comment on possibilities to modify the Bethe permanent so that it approximates the permanent even better, and we conclude the paper with some conjectures about permanent-based pseudo-codewords and permanent-based kernels.

***Index Terms*—Bethe approximation, Bethe permanent, graph cover, partition function, perfect matching, permanent, sum-product algorithm.**

I. Introduction

Central to the topic of this paper is the definition of the permanent of a square matrix (see, *e.g.*, [1]).

Definition 1 *Let $\boldsymbol{\theta} = (\theta_{i,j})_{i,j}$ be a real matrix of size $n \times n$. The permanent of $\boldsymbol{\theta}$ is defined to be the scalar*

$$\mathrm{perm}(\boldsymbol{\theta}) = \sum_{\sigma} \prod_{i \in [n]} \theta_{i,\sigma(i)}, \qquad (1)$$

where the summation is over all $n!$ permutations of the set $[n] \triangleq \{1, 2, \ldots, n\}$. $\square$

Contrast this definition with the definition of the *determinant* of $\boldsymbol{\theta}$, *i.e.*,

$$\det(\boldsymbol{\theta}) = \sum_{\sigma} \mathrm{sgn}(\sigma) \prod_{i \in [n]} \theta_{i,\sigma(i)},$$

where $\mathrm{sgn}(\sigma)$ equals $+1$ if σ is an even permutation and equals -1 if σ is an odd permutation.

A. Complexity of Computing the Permanent

Because the definition of the permanent looks simpler than the definition of the determinant, it is tempting to conclude that the permanent can be computed at least as efficiently as the determinant. However, this does not seem to be the case. Namely, whereas the arithmetic complexity (number of real additions and multiplications) needed to compute the determinant is in $O(n^3)$, Ryser's algorithm (one of the most efficient algorithms for computing the permanent) requires $\Theta(n \cdot 2^n)$ arithmetic operations [2]. This clearly improves upon the brute-force complexity $O(n \cdot n!) = O\big(n^{3/2} \cdot (n/e)^n\big)$ for computing the permanent, but is still exponential in the matrix size.

In terms of complexity classes, the computation of the permanent is in the complexity class #P ("sharp P" or "number P") [3], where #P is the set of the counting problems associated with the decision problems in the class NP. Note that even the computation of the permanent of matrices that contain only zeros and ones is #P-complete. Therefore, the above-mentioned complexity numbers for the computation of the permanent are not surprising.

B. Approximations to the Permanent

Given the difficulty of computing the permanent exactly, and given the fact that in many applications it is good enough to compute an *approximation* to the permanent, this paper focuses on efficient methods to approximate the permanent. This relaxation in requirements, from exact to approximate evaluation of the permanent, allows one to devise algorithms that potentially have much lower complexity.

Moreover, we will consider only the case where the matrix $\boldsymbol{\theta}$ in (1) is non-negative, *i.e.*, where all entries of $\boldsymbol{\theta}$ are non-negative. It is to be expected that approximating the permanent is simpler in this case because with this restriction the sum in (1) contains only non-negative terms, *i.e.*, the terms in this sum "interfere constructively." This is in contrast to the general case where the sum in (1) contains positive and negative terms, *i.e.*, the terms in this sum "interfere constructively and destructively."[1] Despite this restriction to non-negative matrices, many interesting counting problems can be captured by this setup.

Earlier work on approximating the permanent of a non-negative matrix includes Markov-chain-Monte-Carlo-based methods by Broder (see [4]), fully polynomial-time randomized approximation schemes (FPRAS) [5], [6] (for more details, in particular complexity estimates of these methods, see for example the discussion in [6]) and Bethe-approximation-based / sum-product-algorithm (SPA) based methods [7], [8].

Submitted to IEEE Transactions on Information Theory, July 21, 2011. Some of the material in this paper was previously presented at the 48th Annual Allerton Conference on Communications, Control, and Computing, Monticello, IL, USA, Sep. 29–Oct. 1, 2010, and at the 2011 Information Theory and Applications Workshop, UC San Diego, La Jolla, CA, USA, Feb. 6–11 2011.

P. O. Vontobel is with Hewlett–Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA (e-mail: pascal.vontobel@ieee.org).

[1] Strictly speaking, there are also matrices $\boldsymbol{\theta}$ with positive and negative entries but where the product $\prod_{i \in [n]} \theta_{i,\sigma(i)}$ is non-negative for every σ.

The study in this paper was very much motivated by this last set of papers on graphical-model-based methods, in particular by the fact these methods yield algorithms that are *very efficient* and by the fact that the obtained permanent estimates have an *accuracy that is good enough for many purposes*.

The main idea behind this graphical-model-based approach is to formulate a factor graph whose partition function equals the permanent that we are looking for. Consequently, the negative logarithm of the permanent equals the minimum of the so-called Gibbs free energy function that is associated with this factor graph. Although being an elegant reformulation of the permanent computation problem, this does not yet yield any computational savings. Nevertheless, it suggests to look for a function that is tractable and whose minimum is close to the minimum of the Gibbs free energy function. One such function is the so-called Bethe free energy function [9], and with this, paralleling the above-mentioned relationship between the permanent and the minimum of the Gibbs free energy function, the *Bethe permanent* is defined such that its negative logarithm equals the global minimum of the Bethe free energy function. The Bethe free energy function is an interesting candidate because a theorem by Yedidia, Freeman, and Weiss [9] says that fixed points of the SPA correspond to stationary points of the Bethe free energy function.

In general, this approach of replacing the Gibbs free energy function by the Bethe free energy function comes with very few guarantees, though.

- The Bethe free energy function might have multiple local minima.
- It is unclear how close the (global) minimum of the Bethe free energy function is to the minimum of the Gibbs free energy function.
- It is unclear if the SPA converges, even to a local minimum of the Bethe free energy function. (As we will see, the factor graph that we use (*cf.* Figure 1) is not sparse and has many short cycles, in particular many four-cycles. These facts might suggest that the application of the SPA to this factor graph is rather problematic.)

Luckily, in the case of the permanent approximation problem, one can formulate a factor graph where the Bethe free energy function is very well behaved. In particular, in this paper we discuss a factor graph that has the following properties.

- We show that the Bethe free energy function is *convex* and therefore has no non-global local minima.
- The minimum of the Bethe free energy function is *quite close* to the minimum of the Gibbs free energy function. Namely, as was recently shown by Gurvits [10], the permanent is lower bounded by the Bethe permanent. Moreover, we list conjectures on strict and probabilistic Bethe permanent based upper bounds on the permanent. In particular, for certain classes of square non-negative matrices, empirical evidence suggests that the permanent is upper bounded by some constant (that grows rather modestly with the matrix size) times the Bethe permanent.
- We show that the SPA *finds* the minimum of the Bethe free energy function. In fact, the error between the iteration-dependent estimate of the Bethe permanent and the Bethe permanent itself decays exponentially fast, with an exponent depending on the matrix $\boldsymbol{\theta}$. Interestingly enough, in the associated convergence analysis a key role is played by a certain Markov chain that maximizes the sum of its entropy rate plus some average state transition cost.

Besides leaving some questions open with respect to (w.r.t.) the Bethe free energy function (see, *e.g.*, the above-mentioned conjectures concerning permanent upper bounds), these results by-and-large validate the empirical success, as observed by Chertkov *et al.* [7] and by Huang and Jebara [8], of approximating the permanent by graphical-model-based methods.

Let us remark that for many factor graphs with cycles the Bethe free energy function is not as well behaved as the Bethe free energy function under consideration in this paper. In particular, as discussed in [11], every code picked from an ensemble of regular low-density parity-check codes [12], where the ensemble is such that the minimum Hamming distance grows (with high probability) linearly with the block length, has a Bethe free energy function that is concave in certain regions of its domain. Nevertheless, decoding such codes with SPA-based decoders has been highly successful (see, *e.g.* [13]).

C. *Related Work*

The literature on permanents (and adjacent areas of counting perfect matchings, counting zero/one matrices with specified row and column sums, *etc.*.) is vast. Therefore, we just mention works that are (to the best of our knowledge) the most relevant to the present paper.

Besides the already mentioned papers [7], [8] on Bethe-approximation-based methods to the permanent of a non-negative matrix, some aspects of the Bethe free energy function were analyzed by Watanabe and Chertkov in [14] and by Chertkov *et al.* in [15]. (In particular, the paper [14] applied the loop calculus technique by Chertkov and Chernyak [16].) Very recent work in that line of research is presented in a paper by A. B. Yedidia and Chertkov [17] that studies so-called fractional free energy functionals, and resulting lower and upper bounds on the permanent of a non-negative matrix.

Because computing the permanent is related to counting perfect matchings, the paper by Bayati and Nair [18] on counting matchings in graphs with the help of the SPA is very relevant. Note that their setup is such that the perfect matching case can be seen as a limiting case (namely the zero-temperature limit) of the matching setup. However, for the perfect matching case (a case for which the authors of [18] make no claims) the convergence proof of the SPA in [18] is incomplete. Moreover, their matchings are weighted only inasmuch as the weight of a matching depends on the size of the matching. Consequently, because all perfect matchings have the same size, they all are assigned the same weight.

Very relevant to the present paper are also papers on max-product algorithm / min-sum algorithm based approaches to the maximum weight perfect matching problem [19]–[21]. As shown in these papers, these algorithms find the desired

solution efficiently, a fact which is strongly related to the observation that the linear programming relaxation of the underlying integer linear program is tight. This tightness in relaxation, which is an immediate consequence of a theorem by Birkhoff and von Neumann (see Theorem 3), goes also a long way towards explaining why the Bethe free energy function under consideration in this paper is well behaved. Finally, let us remark that because the difference between two perfect matchings corresponds to a union of disjoint cycles, the max-product algorithm / min-sum algorithm convergence analysis in [19]–[21] has some resemblance with Wiberg's max-product algorithm / min-sum algorithm convergence analysis for so-called cycle codes [22].

The present paper has also some similarities with recent papers by Barvinok on counting zero/one matrices with prescribed row and column sums [23] and by Barvinok and Samorodnitsky on computing the partition function for perfect matchings in hypergraphs [24]. However, these papers pursue what would be called a mean-field theory approach in the physics literature [25]. An exception to the previous statement is Section 3.2 in [23], which contains Bethe-approximation-type computations. (See the references in that section for further papers that investigate similar approaches.)

Finally, as already mentioned in the previous subsection, Gurvits's recent paper [10] contains important observations w.r.t. the relationship between the permanent and the Bethe permanent of a non-negative matrix, and puts them into the context of Schrijver's permanental inequality.

D. Overview of the Paper

This paper is structured as follows. We conclude this introductory section with a discussion of some of the notation that is used. In Section II we then introduce the main normal factor graph (NFG) for this paper, in Section III we formally define the Bethe permanent, in Section IV we discuss properties of the Bethe entropy function and the Bethe free energy function, in Section V we analyze the SPA, in Section VI we give a "combinatorial characterization" of the Bethe permanent in terms of graph covers of the above-mentioned NFG, in Section VII we discuss Bethe-permanent-based bounds on the permanent, in Section VIII we list some thoughts on using the concept of the "fractional Bethe entropy function," in Section IX we list some conjectures, and we conclude the paper in Section X. Finally, the appendix contains some of the proofs.

E. Basic Notations and Definitions

This subsection discusses the most important notations that will be used in this paper. More notational definitions will be given in later sections.

We let $\mathbb{R}$ be the field of real numbers, $\mathbb{R}_{\geqslant 0}$ be the set of non-negative real numbers, $\mathbb{R}_{>0}$ be the set of positive real number, $\mathbb{Z}$ be the ring of integers, $\mathbb{Z}_{\geqslant 0}$ be the set of non-negative integers, and $\mathbb{Z}_{>0}$ be the set of positive integers. Scalars are denoted by non-boldface characters, whereas vectors and matrices by boldface characters. For any positive integer L, the matrix $\mathbf{1}_{L\times L}$ is the all-one matrix of size $L \times L$.

Assumption 2 *Throughout this paper, if not mentioned otherwise, n is a positive integer and $\boldsymbol{\theta} = (\theta_{i,j})_{i,j}$ is a non-negative matrix of size $n \times n$.* □

We use calligraphic letters for sets, and the size of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. The convex hull [26] of some subset $\mathcal{S}$ of some multi-dimensional real space is denoted by $\mathrm{conv}(\mathcal{S})$. For any positive integer L we define $[L] \triangleq \{1, \ldots, L\}$. For any positive integer L, we define $\mathcal{P}_{L\times L}$ to be the set of all $L \times L$ permutation matrices, *i.e.*,

$$\mathcal{P}_{L\times L} \triangleq \left\{ \boldsymbol{P} \;\middle|\; \begin{array}{l} \boldsymbol{P} \text{ is a matrix of size } L \times L \\ \boldsymbol{P} \text{ contains exactly one } 1 \text{ per row} \\ \boldsymbol{P} \text{ contains exactly one } 1 \text{ per column} \\ \boldsymbol{P} \text{ contains } 0\text{s otherwise} \end{array} \right\},$$

Clearly, there is a bijection between $\mathcal{P}_{L\times L}$ and the set of all permutations of $[L]$. Moreover, for a finite set $\mathcal{S}$, we define $\Pi_{\mathcal{S}}$ to be the set of probability mass functions over $\mathcal{S}$, *i.e.*,

$$\Pi_{\mathcal{S}} \triangleq \left\{ \boldsymbol{p} = (p_s)_{s\in\mathcal{S}} \;\middle|\; p_s \geqslant 0 \text{ for all } s \in \mathcal{S},\ \sum_{s\in\mathcal{S}} p_s = 1 \right\}.$$

Finally, for any positive integer L, we let $\Gamma_{L\times L}$ be the set of doubly stochastic matrices of size $L \times L$, *i.e.*,

$$\Gamma_{L\times L} \triangleq \left\{ \boldsymbol{\gamma} = (\gamma_{i,j}) \;\middle|\; \begin{array}{l} \gamma_{i,j} \geqslant 0 \text{ for all } (i,j) \in [L] \times [L] \\ \sum_{j\in[L]} \gamma_{i,j} = 1 \text{ for all } i \in [L] \\ \sum_{i\in[L]} \gamma_{i,j} = 1 \text{ for all } j \in [L] \end{array} \right\}.$$

In the rest of the paper, when appropriate, we will identify the set of $L \times L$ real matrices with the L^2-dimensional real space. In that sense, $\Gamma_{L\times L}$ can be seen as a polytope in the L^2-dimensional real space. Clearly, $\Gamma_{L\times L}$ is a convex set, and every permutation matrix of size $L \times L$ is a doubly stochastic matrix of size $L \times L$. Most interestingly, every doubly stochastic matrix of size $L \times L$ can be written as a convex combination of permutation matrices of size $L \times L$; this observation is a consequence of the important Birkhoff–von Neumann Theorem.

Theorem 3 (Birkhoff–von Neumann Theorem) *For any positive integer L, the set of doubly stochastic matrices of size $L \times L$ is a polytope whose vertex set equals the set of permutation matrices of size $L \times L$,* i.e.,

$$\text{vertex-set}(\Gamma_{L\times L}) = \mathcal{P}_{L\times L}.$$

As a consequence, the set of doubly stochastic matrices of size $L \times L$ is the convex hull of the set of all permutation matrices of size $L \times L$, i.e.,

$$\Gamma_{L\times L} = \mathrm{conv}(\mathcal{P}_{L\times L}).$$

Proof: See, *e.g.*, [27, Section 8.7]. □

Finally, all logarithms will be natural logarithms and the value of $0 \cdot \log(0)$ is defined to be equal to 0.

II. Normal Factor Graph Representation

Factor graphs are a convenient way to represent multivariate functions [28]. In this paper we use a variant called "normal factor graphs (NFGs)" [29] (also called "Forney-style factor graphs" [30]), where variables are associated with edges.

As already mentioned in the introduction, the main idea behind the graphical-model-based approach to estimating the permanent is to formulate an NFG such that its partition function equals the permanent. There are of course different ways to do this and typically different formulations will yield different results when estimating the permanent with sub-optimal algorithms like the SPA. It is well known that when the NFG has no cycles, then the SPA computes the partition function exactly, however, for the given problem any NFG *without* cycles yields highly inefficient SPA update rules for reasonably large n (otherwise there would be a contradiction to the considerations in Section I-A), and so we will focus on NFGs *with* cycles. The NFG that is introduced in the following definition and that is based on a complete bipartite graph with two times n vertices, is a rather natural candidate, and, as we will see, has very interesting and useful properties.

Definition 4 *We define the NFG* $\mathsf{N}(\boldsymbol{\theta}) \triangleq \mathsf{N}(\mathcal{F}, \mathcal{E}, \mathcal{A}, \mathcal{G})$ *as follows (see also Figure 1).*

- *The set of vertices is* $\mathcal{F} \triangleq \mathcal{I} \,\dot{\cup}\, \mathcal{J}$, *where* $\mathcal{I} \triangleq [n]$ *will be called the set of left vertices and* $\mathcal{J} \triangleq [n]$ *will be called the set of right vertices.*[2]
- *The set of edges is* $\mathcal{E} \triangleq \mathcal{I} \times \mathcal{J} = \big\{(i,j) \,\big|\, i \in \mathcal{I}, j \in \mathcal{J}\big\}$. *Moreover, all edges are full edges,* i.e., $\mathcal{E}_{\mathrm{full}} = \mathcal{E}$ *and* $\mathcal{E}_{\mathrm{half}} = \{\,\}$.
- *With every edge* $e = (i,j) \in \mathcal{E}$ *we associate the variable* $A_e = A_{i,j}$ *with alphabet* $\mathcal{A}_e = \mathcal{A}_{i,j} \triangleq \{0,1\}$; *a realization of* $A_e = A_{i,j}$ *will be denoted by* $a_e = a_{i,j}$.
- *The set* $\mathcal{A} \triangleq \prod_e \mathcal{A}_e = \prod_{i,j} \mathcal{A}_{i,j}$ *will be called the configuration set, and so*

$$\boldsymbol{a} \;\triangleq\; (a_e)_{e \in \mathcal{E}} = (a_{i,j})_{(i,j) \in \mathcal{I} \times \mathcal{J}} \;\in\; \mathcal{A}$$

  will be called a configuration. For a given vector $\boldsymbol{a}$, *we also define the sub-vectors*

$$\boldsymbol{a}_i \triangleq \big\{(a_{i,j})_{j \in \mathcal{J}}\big\} \quad \textit{and} \quad \boldsymbol{a}_j \triangleq \big\{(a_{i,j})_{i \in \mathcal{I}}\big\}.$$

  When convenient, the vector $\boldsymbol{a}$ *will be considered to be an* $n \times n$ *matrix. Then* $\boldsymbol{a}_i$ *corresponds to the* i*-th row of* $\boldsymbol{a}$, *and* $\boldsymbol{a}_j$ *corresponds to the* j*-th column of* $\boldsymbol{a}$.
- *For every* $i \in \mathcal{I}$ *we define the local functions*[3] [4]

$$g_i : \prod_{j'} \mathcal{A}_{i,j'} \to \mathbb{R}, \quad \boldsymbol{a}_i \mapsto \begin{cases} \sqrt{\theta_{i,j}} & \textit{(if } \boldsymbol{a}_i = \boldsymbol{u}_j\textit{)} \\ 0 & \textit{(otherwise)} \end{cases}$$

[2]Here, $\mathcal{F} \triangleq \mathcal{I} \,\dot{\cup}\, \mathcal{J}$ stands for the more cumbersome $\mathcal{F} \triangleq \big(\{\mathrm{left}\} \times \mathcal{I}\big) \cup \big(\{\mathrm{right}\} \times \mathcal{J}\big)$. In the following, i (and variations thereof) will refer to a left vertex and j (and variations thereof) will refer to a right vertex. In that spirit, variables like η_i and η_j are different variables, also if $i = j$.

[3]Here and in the following, $\boldsymbol{u}_j$, $j \in \mathcal{J}$, stands for the length-n vector where all entries are zero except for the j-th entry that equals 1. The vector $\boldsymbol{u}_i$, $i \in \mathcal{I}$, is defined similarly.

[4]Here and in the following, we will use the short-hands $\sum_i$, $\sum_j$, $\sum_{i'}$, $\sum_{j'}$, $\sum_e$, $\sum_{e'}$ for $\sum_{i \in \mathcal{I}}$, $\sum_{j \in \mathcal{J}}$, $\sum_{i' \in \mathcal{I}}$, $\sum_{j' \in \mathcal{J}}$, $\sum_{e \in \mathcal{E}}$, $\sum_{e' \in \mathcal{E}}$, respectively, with similar conventions for products.

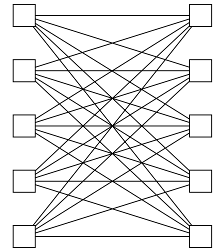


Fig. 1. The NFG $\mathsf{N}(\boldsymbol{\theta})$ which is based on a complete bipartite graph with two times n vertices (here $n = 5$). The left function nodes represent the functions $\{g_i\}_{i \in \mathcal{I}}$, the right function nodes represent the functions $\{g_j\}_{j \in \mathcal{J}}$, and with the edge $e = (i,j)$ we associate the variable $A_e = A_{i,j}$. (See Definition 4 for more details.)

  Similarly, for every $j \in \mathcal{J}$ *we define the local functions*

$$g_j : \prod_{i'} \mathcal{A}_{i',j} \to \mathbb{R}, \quad \boldsymbol{a}_j \mapsto \begin{cases} \sqrt{\theta_{i,j}} & \textit{(if } \boldsymbol{a}_j = \boldsymbol{u}_i\textit{)} \\ 0 & \textit{(otherwise)} \end{cases}$$

- *For every* $i \in \mathcal{I}$ *we define the function node alphabet* $\mathcal{A}_i$ *to be the set*

$$\mathcal{A}_i \triangleq \left\{ \boldsymbol{a}_i \in \prod_{j'} \mathcal{A}_{i,j'} \;\middle|\; g_i(\boldsymbol{a}_i) \neq 0 \right\} = \{\boldsymbol{u}_j \mid j \in \mathcal{J}\}.$$

  Similarly, for every $j \in \mathcal{J}$ *we define the function node alphabet* $\mathcal{A}_j$ *to be the set*

$$\mathcal{A}_j \triangleq \left\{ \boldsymbol{a}_j \in \prod_{i'} \mathcal{A}_{i',j} \;\middle|\; g_j(\boldsymbol{a}_j) \neq 0 \right\} = \{\boldsymbol{u}_i \mid i \in \mathcal{I}\}.$$

  (The sets $\mathcal{A}_i$ *and* $\mathcal{A}_j$ *are also known as local constraint code of the function nodes* i *and* j*, respectively.)*
- *The global function* g *is defined to be*

$$g : \mathcal{A} \to \mathbb{R}, \quad \boldsymbol{a} \mapsto \left(\prod_i g_i(\boldsymbol{a}_i) \right) \cdot \left(\prod_j g_j(\boldsymbol{a}_j) \right).$$

- *A configuration* $\boldsymbol{c}$ *with* $g(\boldsymbol{c}) \neq 0$ *will be called a valid configuration. The set of all valid configurations,* i.e.,

$$\mathcal{C}_{\mathcal{E}} \triangleq \left\{ (c_{i,j})_{i,j \in \mathcal{I} \times \mathcal{J}} \;\middle|\; \begin{array}{c} c_{i,j} \in \mathcal{A}_{i,j},\ (i,j) \in \mathcal{I} \times \mathcal{J} \\ \boldsymbol{c}_i \in \mathcal{A}_i,\ i \in \mathcal{I} \\ \boldsymbol{c}_j \in \mathcal{A}_j,\ j \in \mathcal{J} \end{array} \right\};$$

  will be called the global behavior of $\mathsf{N}(\boldsymbol{\theta})$*. Considering the elements of* $\mathcal{C}_{\mathcal{E}}$ *as* $n \times n$ *matrices, it can easily be verified that* $\mathcal{C}_{\mathcal{E}} = \mathcal{P}_{n \times n}$*. This allows us to associate with* $\boldsymbol{c} \in \mathcal{C}_{\mathcal{E}}$ *the permutation* $\sigma_{\boldsymbol{c}} : [n] \to [n]$ *that maps* $i \in \mathcal{I}$ *to* $j \in \mathcal{J}$ *if* $c_{i,j} = 1$. □

Lemma 5 *Consider the NFG* $\mathsf{N}(\boldsymbol{\theta})$ *and let* $\boldsymbol{c} \in \mathcal{C}_{\mathcal{E}}$ *be a valid configuration of it. Then*

$$g_i(\boldsymbol{c}_i) = \sqrt{\theta_{i,\sigma_{\boldsymbol{c}}(i)}}, \quad i \in \mathcal{I},$$
$$g_j(\boldsymbol{c}_j) = \sqrt{\theta_{\sigma_{\boldsymbol{c}}^{-1}(j),j}}, \quad j \in \mathcal{J},$$
$$g(\boldsymbol{c}) = \prod_i \theta_{i,\sigma_{\boldsymbol{c}}(i)} = \prod_j \theta_{\sigma_{\boldsymbol{c}}^{-1}(j),j}.$$

Proof: The first two expressions follow easily from the definitions of g_i and g_j in Definition 4. The third expression follows from

$$\begin{aligned} g(\boldsymbol{c}) &= \left(\prod_i g_i(\boldsymbol{c}_i)\right) \cdot \left(\prod_j g_j(\boldsymbol{c}_j)\right) \\ &= \left(\prod_i \sqrt{\theta_{i,\sigma_{\boldsymbol{c}}(i)}}\right) \cdot \left(\prod_j \sqrt{\theta_{\sigma_{\boldsymbol{c}}^{-1}(j),j}}\right) \\ &= \left(\prod_i \sqrt{\theta_{i,\sigma_{\boldsymbol{c}}(i)}}\right) \cdot \left(\prod_{i'} \sqrt{\theta_{i',\sigma_{\boldsymbol{c}}(i')}}\right) \\ &= \prod_i \theta_{i,\sigma_{\boldsymbol{c}}(i)}. \end{aligned}$$

□

Definition 6 *The (Gibbs) partition function of the NFG* $\mathsf{N}(\boldsymbol{\theta})$ *is defined to be the sum of the global function over all configurations, or, equivalently, the sum of the global function over all valid configurations,* i.e.,

$$Z_{\mathrm{G}} \triangleq \sum_{\boldsymbol{a} \in \mathcal{A}} g(\boldsymbol{a}) = \sum_{\boldsymbol{c} \in \mathcal{C}_{\mathcal{E}}} g(\boldsymbol{c}). \tag{2}$$

In the following, when confusion can arise what NFG a certain Gibbs partition function is referring to, we will use $Z_{\mathrm{G}}\big(\mathsf{N}(\boldsymbol{\theta})\big)$, *etc., instead of* Z_{G}.[5] □

Definition 7 *The Gibbs free energy function associated with the NFG* $\mathsf{N}(\boldsymbol{\theta})$ *is defined to be*

$$F_{\mathrm{G}} : \Pi_{\mathcal{C}_{\mathcal{E}}} \to \mathbb{R}, \quad \boldsymbol{p} \mapsto U_{\mathrm{G}}(\boldsymbol{p}) - H_{\mathrm{G}}(\boldsymbol{p}),$$

where

$$\begin{aligned} U_{\mathrm{G}} &: \Pi_{\mathcal{C}_{\mathcal{E}}} \to \mathbb{R}, \quad \boldsymbol{p} \mapsto -\sum_{\boldsymbol{c} \in \mathcal{C}_{\mathcal{E}}} p_{\boldsymbol{c}} \cdot \log\big(g(\boldsymbol{c})\big), \\ H_{\mathrm{G}} &: \Pi_{\mathcal{C}_{\mathcal{E}}} \to \mathbb{R}, \quad \boldsymbol{p} \mapsto -\sum_{\boldsymbol{c} \in \mathcal{C}_{\mathcal{E}}} p_{\boldsymbol{c}} \cdot \log\big(p_{\boldsymbol{c}}\big). \end{aligned}$$

Here, U_{G} *is called the Gibbs average energy function and* H_{G} *is called the Gibbs entropy function. In the following, when confusion can arise what NFG a certain Gibbs free energy function is referring to, we will use* $F_{\mathrm{G},\mathsf{N}(\boldsymbol{\theta})}$, *etc., instead of* F_{G}. *Similar comments apply to* U_{G} *and* H_{G}. □

For more details on these functions we refer to, *e.g.*, [9]. For a discussion of these functions in the context of NFGs we refer to, *e.g.*, [11]. Note that H_{G} is a concave function of $\boldsymbol{p}$, that U_{G} is a linear function of $\boldsymbol{p}$, and that, consequently, F_{G} is a convex function of $\boldsymbol{p}$.

[5]Note that "function" in "partition function" refers to the fact that the expression in (2) typically is a function of some parameters like the temperature T (see the discussion below). A better word for "partition function" would possibly be "partition sum" or "state sum," which would more closely follow the German "Zustandssumme" whose first letter is used to denote the partition function.

Lemma 8 *The permanent of* $\boldsymbol{\theta}$ *can be expressed in terms of the Gibbs partition function or in terms of the minimum of the Gibbs free energy function of* $\mathsf{N}(\boldsymbol{\theta})$. *Namely,*

$$\mathrm{perm}(\boldsymbol{\theta}) = Z_{\mathrm{G}} = \exp\left(-\min_{\boldsymbol{p}} F_{\mathrm{G}}(\boldsymbol{p})\right), \tag{3}$$

where the minimization is over $\boldsymbol{p} \in \Pi_{\mathcal{C}_{\mathcal{E}}}$.

Proof: The first equality is a straightforward consequence of Definitions 1 and 4, along with Lemma 5. For the second equality we refer to, *e.g.*, [9], [11]. □

The Gibbs partition function Z_{G} and the Gibbs free energy function F_{G} were specified for temperature $T = 1$ in the above definitions. For a general temperature parameter $T \in \mathbb{R}_{\geqslant 0}$, these functions have to be replaced by $Z_{\mathrm{G}} \triangleq \sum_{\boldsymbol{c} \in \mathcal{C}_{\mathcal{E}}} g(\boldsymbol{c})^{1/T}$ and by $F_{\mathrm{G}}(\boldsymbol{p}) \triangleq U_{\mathrm{G}}(\boldsymbol{p}) - T \cdot H_{\mathrm{G}}(\boldsymbol{p})$, respectively, and Lemma 8 has to be replaced by $Z_{\mathrm{G}} = \exp\big(-\frac{1}{T}\min_{\boldsymbol{p}} F_{\mathrm{G}}(\boldsymbol{p})\big)$. Of course, $Z_{\mathrm{G}} = \mathrm{perm}(\boldsymbol{\theta})$ does not hold anymore, unless a suitable T-dependence is built into the definition of $\mathrm{perm}(\boldsymbol{\theta})$.

III. The Bethe Permanent

Although the reformulation of the permanent in the above lemma in terms of a convex minimization problem is elegant, from a computational perspective it does not buy us much. However, it suggest to look for a minimization problem that can be solved efficiently and whose minimal value is related to the desired quantity. This is the approach that is taken in this section and will be based on the Bethe approximation of the Gibbs free energy function: the resulting approximation of the permanent of a non-negative square matrix will be called the Bethe permanent. (Note that in this section we give the technical details only; for a general discussion w.r.t. the motivations behind the Bethe approximation we refer to [9], and for a discussion of the Bethe approximation in the context of NFGs we refer to [11].)

Definition 9 *Consider the NFG* $\mathsf{N}(\boldsymbol{\theta})$. *We let*

$$\boldsymbol{\beta} \triangleq \big((\boldsymbol{\beta}_i)_{i \in \mathcal{I}}, (\boldsymbol{\beta}_j)_{j \in \mathcal{J}}, (\boldsymbol{\beta}_e)_{e \in \mathcal{E}}\big)$$

be a collection of vectors based on the real vectors

$$\begin{aligned} \boldsymbol{\beta}_i &\triangleq (\beta_{i,\boldsymbol{a}_i})_{\boldsymbol{a}_i \in \mathcal{A}_i}, \\ \boldsymbol{\beta}_j &\triangleq (\beta_{j,\boldsymbol{a}_j})_{\boldsymbol{a}_j \in \mathcal{A}_j}, \\ \boldsymbol{\beta}_e &\triangleq (\beta_{e,a_e})_{a_e \in \mathcal{A}_e}. \end{aligned}$$

Moreover, we define the sets

$$\begin{aligned} \mathcal{B}_i &\triangleq \Pi_{\mathcal{A}_i}, \quad i \in \mathcal{I}, \\ \mathcal{B}_j &\triangleq \Pi_{\mathcal{A}_j}, \quad j \in \mathcal{J}, \\ \mathcal{B}_e &\triangleq \Pi_{\mathcal{A}_e}, \quad e \in \mathcal{E}, \end{aligned}$$

and call $\mathcal{B}_i$, $\mathcal{B}_j$, *and* $\mathcal{B}_e$, *the* i*th local marginal polytope, the* j*th local marginal polytope, the* e*th local marginal polytope, respectively. (Sometimes* $\mathcal{B}_i$ *is also called the* i*th belief polytope, etc.)*

With this, the local marginal polytope (or belief polytope) $\mathcal{B}$ is defined to be the set

$$\mathcal{B} = \left\{ \boldsymbol{\beta} \;\middle|\; \begin{array}{c} \boldsymbol{\beta}_i \in \mathcal{B}_i \text{ for all } i \in \mathcal{I} \\ \boldsymbol{\beta}_j \in \mathcal{B}_j \text{ for all } j \in \mathcal{J} \\ \boldsymbol{\beta}_e \in \mathcal{B}_e \text{ for all } e \in \mathcal{E} \\ \displaystyle\sum_{\boldsymbol{a}'_i \in \mathcal{A}_i :\, a'_{i,e} = a_e} \beta_{i,\boldsymbol{a}'_i} = \beta_{e,a_e} \\ \text{for all } e = (i,j) \in \mathcal{E},\ a_e \in \mathcal{A}_e \\ \displaystyle\sum_{\boldsymbol{a}'_j \in \mathcal{A}_j :\, a'_{j,e} = a_e} \beta_{j,\boldsymbol{a}'_j} = \beta_{e,a_e} \\ \text{for all } e = (i,j) \in \mathcal{E},\ a_e \in \mathcal{A}_e \end{array} \right\},$$

where $\boldsymbol{\beta} \in \mathcal{B}$ is called a pseudo-marginal vector. (The two constraints that were listed last in the definition of $\mathcal{B}$ will be called "edge consistency constraints.") □

Definition 10 *The Bethe free energy function associated with the NFG $\mathsf{N}(\boldsymbol{\theta})$ is defined to be the function*

$$F_{\mathrm{B}} : \mathcal{B} \to \mathbb{R}, \quad \boldsymbol{\beta} \mapsto U_{\mathrm{B}}(\boldsymbol{\beta}) - H_{\mathrm{B}}(\boldsymbol{\beta}),$$

where

$$\begin{aligned} U_{\mathrm{B}} : \mathcal{B} \to \mathbb{R}, \quad & \boldsymbol{\beta} \mapsto \sum_i U_{\mathrm{B},i}(\boldsymbol{\beta}_i) + \sum_j U_{\mathrm{B},j}(\boldsymbol{\beta}_j) \\ H_{\mathrm{B}} : \mathcal{B} \to \mathbb{R}, \quad & \boldsymbol{\beta} \mapsto \sum_i H_{\mathrm{B},i}(\boldsymbol{\beta}_i) + \sum_j H_{\mathrm{B},j}(\boldsymbol{\beta}_j) \\ & \qquad - \sum_e H_{\mathrm{B},e}(\boldsymbol{\beta}_e), \end{aligned}$$

with[6]

$$\begin{aligned} U_{\mathrm{B},i} : \mathcal{B}_i \to \mathbb{R}, \quad & \boldsymbol{\beta}_i \mapsto -\sum_{\boldsymbol{a}_i} \beta_{i,\boldsymbol{a}_i} \cdot \log\big(g_i(\boldsymbol{a}_i)\big), \\ U_{\mathrm{B},j} : \mathcal{B}_j \to \mathbb{R}, \quad & \boldsymbol{\beta}_j \mapsto -\sum_{\boldsymbol{a}_j} \beta_{j,\boldsymbol{a}_j} \cdot \log\big(g_j(\boldsymbol{a}_j)\big), \\ H_{\mathrm{B},i} : \mathcal{B}_i \to \mathbb{R}, \quad & \boldsymbol{\beta}_i \mapsto -\sum_{\boldsymbol{a}_i} \beta_{i,\boldsymbol{a}_i} \cdot \log\big(\beta_{i,\boldsymbol{a}_i}\big), \\ H_{\mathrm{B},j} : \mathcal{B}_j \to \mathbb{R}, \quad & \boldsymbol{\beta}_j \mapsto -\sum_{\boldsymbol{a}_j} \beta_{j,\boldsymbol{a}_j} \cdot \log\big(\beta_{j,\boldsymbol{a}_j}\big), \\ H_{\mathrm{B},e} : \mathcal{B}_e \to \mathbb{R}, \quad & \boldsymbol{\beta}_e \mapsto -\sum_{a_e} \beta_{e,a_e} \cdot \log\big(\beta_{e,a_e}\big). \end{aligned}$$

Here, U_{B} is the Bethe average energy function and H_{B} is the Bethe entropy function. In the following, when confusion can arise what NFG a certain Bethe free energy function is referring to, we will use $F_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}$, etc., instead of F_{B}. Similar comments apply to U_{B} and H_{B}. □

With this, the Bethe partition function of an NFG is *defined* such that an equality analogous to the second equality in (3) holds.

Definition 11 *The Bethe partition function of the NFG $\mathsf{N}(\boldsymbol{\theta})$ is defined to be*

$$Z_{\mathrm{B}} \triangleq \exp\left(- \min_{\boldsymbol{\beta} \in \mathcal{B}} F_{\mathrm{B}}(\boldsymbol{\beta}) \right).$$

In the following, when confusion can arise what NFG a certain Bethe partition function is referring to, we will use $Z_{\mathrm{B}}(\mathsf{N})$, etc., instead of Z_{B}. □

The next definition is the main definition of this paper and was motivated by the work of Chertkov *et al.* [7] and by the work of Huang and Jebara [8].

Definition 12 *Consider the NFG $\mathsf{N}(\boldsymbol{\theta})$. The Bethe permanent of $\boldsymbol{\theta}$, which will be denoted by $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})$, is defined to be*

$$\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) \triangleq Z_{\mathrm{B}}\big(\mathsf{N}(\boldsymbol{\theta})\big).$$

□

A similar comment w.r.t. a temperature parameter $T \in \mathbb{R}_{\geqslant 0}$ as at the end of Section II applies also to the definition of the Bethe partition function and the Bethe free energy function. In the following, however, we will only consider the case $T = 1$. An exception is Section VIII on the fractional Bethe approximation: this approximation can be viewed as introducing multiple temperature parameters, namely one temperature parameter for every term of H_{B}, and therefore includes the single temperature parameter case as a special case.

IV. Properties of the Bethe Entropy Function and the Bethe Free Energy Function

There are relatively few general statements about the shape of the Bethe entropy function. In this section we show that Bethe entropy function associated with $\mathsf{N}(\boldsymbol{\theta})$ has many special properties.

- In general, the Bethe entropy function is not a concave function. However, here we show that the Bethe entropy function under consideration, when suitably parametrized, is a concave function.
  Similarly, the Bethe free energy function is in general not a convex function. However, because the Bethe free energy function is the difference of the Bethe average energy function and the Bethe entropy function, because the Bethe average energy function is linear in its arguments, and because the Bethe entropy function is concave, the Bethe free energy function under consideration is convex and does *not* have non-global local minima.[7]
- In general, the Bethe entropy function can take on positive, zero, and negative values. However, here we show that the Bethe entropy under consideration is non-negative.
- Very often, the directional derivative of the Bethe entropy function away from a vertex of its domain is $+\infty$ or $-\infty$. Here we show that the directional derivative of the Bethe entropy function under consideration has a (non-negative) finite slope away from any vertex of its domain. (As we will see in Section V, this observation will have important consequences for the SPA convergence analysis.)

[6] Here and in the following, we use the short-hand $\sum_{\boldsymbol{a}_i}$ for $\sum_{\boldsymbol{a}_i \in \mathcal{A}_i}$, *etc.*.

[7] The fact that convexity / non-convexity of a function depends on its parametrization might explain the non-convexity observations in [8, Section 3.3] w.r.t. the Bethe free energy function.

A. Reformulation of the Bethe Entropy Function and the Bethe Free Energy Function

As mentioned in Section I-C, the successes of the max-product algorithm / min-sum algorithm based approaches to the maximum weight perfect matching problem in the papers [19]–[21] was heavily based on a theorem by Birkhoff and von Neumann (*cf.* Theorem 3). This theorem is equally central to the results of the present paper. Namely, in the next lemma we introduce a parametrization of the belief polytope $\mathcal{B}$ based on $\Gamma_{n\times n}$ that will be used for the rest of the paper.

Lemma 13 *Consider the NFG $\mathsf{N}(\boldsymbol{\theta})$. Its belief polytope $\mathcal{B}$ can be parametrized by $\Gamma_{n\times n}$, the set of doubly stochastic matrices of size $n \times n$. In particular, we define the parametrization such that the matrix $\boldsymbol{\gamma} = (\gamma_{i,j})_{(i,j)\in\mathcal{I}\times\mathcal{J}} \in \Gamma_{n\times n}$ indexes the pseudo-marginal vector $\boldsymbol{\beta} \in \mathcal{B}$ with*

$$\beta_{i,\boldsymbol{a}_i}\Big|_{\boldsymbol{a}_i=\boldsymbol{u}_j} = \beta_{j,\boldsymbol{a}_j}\Big|_{\boldsymbol{a}_j=\boldsymbol{u}_i} = \gamma_{i,j},$$

and

$$\beta_{e,a_e}\Big|_{a_e=0} = 1-\gamma_{i,j}, \quad \beta_{e,a_e}\Big|_{a_e=1} = \gamma_{i,j},$$

for every $i \in \mathcal{I}$, $j \in \mathcal{J}$, and $e = (i,j) \in \mathcal{E}$.

Proof: It is straightforward to verify that the pseudo-marginal vector $\boldsymbol{\beta}$ which is specified in the lemma statement is indeed in $\mathcal{B}$. Moreover, with the help of Theorem 3 one can verify that for every pseudo-marginal vector $\boldsymbol{\beta} \in \mathcal{B}$ there is a $\boldsymbol{\gamma} \in \Gamma_{n\times n}$ such that $\boldsymbol{\gamma}$ indexes $\boldsymbol{\beta}$. □

In the following, for a given matrix $\boldsymbol{\gamma} = (\gamma_{i,j})_{(i,j)\in\mathcal{I}\times\mathcal{J}}$, the i-th row of $\boldsymbol{\gamma}$ will be denoted by $\boldsymbol{\gamma}_i = (\gamma_{i,j})_{j\in\mathcal{J}}$ and the j-th column of $\boldsymbol{\gamma}$ will be denoted by $\boldsymbol{\gamma}_j = (\gamma_{i,j})_{i\in\mathcal{I}}$.

The above observations allow us to express the Bethe free energy function and related functions in terms of $\boldsymbol{\gamma} \in \Gamma_{n\times n}$.

Lemma 14 *Consider the NFG $\mathsf{N}(\boldsymbol{\theta})$. Then*

$$F_{\mathrm{B}} : \Gamma_{n\times n} \to \mathbb{R}, \quad \boldsymbol{\gamma} \mapsto U_{\mathrm{B}}(\boldsymbol{\gamma}) - H_{\mathrm{B}}(\boldsymbol{\gamma}),$$

where

$$\begin{aligned} U_{\mathrm{B}} : \Gamma_{n\times n} \to \mathbb{R}, \quad & \boldsymbol{\gamma} \mapsto \sum_i U_{\mathrm{B},i}(\boldsymbol{\gamma}_i) + \sum_j U_{\mathrm{B},j}(\boldsymbol{\gamma}_j), \\ H_{\mathrm{B}} : \Gamma_{n\times n} \to \mathbb{R}, \quad & \boldsymbol{\gamma} \mapsto \sum_i H_{\mathrm{B},i}(\boldsymbol{\gamma}_i) + \sum_j H_{\mathrm{B},j}(\boldsymbol{\gamma}_i) \\ & \qquad - \sum_{i,j} H_{\mathrm{B},(i,j)}(\gamma_{i,j}), \end{aligned}$$

with

$$\begin{aligned} U_{\mathrm{B},i} :\ \Pi_{[n]} \to \mathbb{R}, \quad & \boldsymbol{\gamma}_i \mapsto -\frac{1}{2}\sum_j \gamma_{i,j}\cdot\log(\theta_{i,j}), \\ U_{\mathrm{B},j} :\ \Pi_{[n]} \to \mathbb{R}, \quad & \boldsymbol{\gamma}_j \mapsto -\frac{1}{2}\sum_i \gamma_{i,j}\cdot\log(\theta_{i,j}), \\ H_{\mathrm{B},i} :\ \Pi_{[n]} \to \mathbb{R}, \quad & \boldsymbol{\gamma}_i \mapsto -\sum_j \gamma_{i,j}\cdot\log(\gamma_{i,j}), \\ H_{\mathrm{B},j} :\ \Pi_{[n]} \to \mathbb{R}, \quad & \boldsymbol{\gamma}_j \mapsto -\sum_i \gamma_{i,j}\cdot\log(\gamma_{i,j}), \end{aligned}$$

$$\begin{aligned} H_{\mathrm{B},(i,j)} :\ & [0,1] \to \mathbb{R} \\ & \gamma_{i,j} \mapsto -\gamma_{i,j}\log(\gamma_{i,j}) - (1-\gamma_{i,j})\log(1-\gamma_{i,j}), \end{aligned}$$

Proof: This follows straightforwardly from Definition 10 and Lemma 13. □

Corollary 15 *It holds that*

$$\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) = \exp\left(-\min_{\boldsymbol{\gamma}\in\Gamma_{n\times n}} F_{\mathrm{B}}(\boldsymbol{\gamma})\right),$$

where

$$\begin{aligned} F_{\mathrm{B}}(\boldsymbol{\gamma}) &= U_{\mathrm{B}}(\boldsymbol{\gamma}) - H_{\mathrm{B}}(\boldsymbol{\gamma}), \\ U_{\mathrm{B}}(\boldsymbol{\gamma}) &= -\sum_{i,j} \gamma_{i,j}\log(\theta_{i,j}), \\ H_{\mathrm{B}}(\boldsymbol{\gamma}) &= -\sum_{i,j} \gamma_{i,j}\log(\gamma_{i,j}) + \sum_{i,j}(1-\gamma_{i,j})\log(1-\gamma_{i,j}). \end{aligned}$$

Proof: This follows from Definitions 11 and 12 and from Lemma 14. □

If the sign in front of the second half of the expression for $H_{\mathrm{B}}(\boldsymbol{\gamma})$ in Corollary 15 were a minus sign, then $H_{\mathrm{B}}(\boldsymbol{\gamma})$ could be expressed as a sum of binary entropy functions, and therefore the concavity of $H_{\mathrm{B}}(\boldsymbol{\gamma})$ would be immediate. However, the presence of the plus sign means that a more careful look at $H_{\mathrm{B}}(\boldsymbol{\gamma})$ is required to determine if it is concave or not.

Assumption 16 *For the rest of this section we assume that $n \geqslant 2$ and that $\boldsymbol{\theta}$ is a* positive *matrix of size $n \times n$. This simplifies the wording of most results without hurting their generality too much. In practice, two possible ways to deal with the issue of zero entries in $\boldsymbol{\theta}$ are the following.*

- *One can change the matrix $\boldsymbol{\theta}$ so that zero entries become tiny positive entries.*
- *One can redefine $\mathsf{N}(\boldsymbol{\theta})$ by removing the edge $e = (i,j)$, along with redefining the local functions g_i and g_j, if $\theta_{i,j} = 0$* □

B. Concavity of the Bethe Entropy Function and Convexity of the Bethe Free Energy Function

Towards showing that $H_{\mathrm{B}}(\boldsymbol{\gamma})$ is a concave function of $\boldsymbol{\gamma}$, and subsequently that $F_{\mathrm{B}}(\boldsymbol{\gamma})$ is a convex function of $\boldsymbol{\gamma}$, we first study two useful functions. Namely, in Definition 17 and Lemma 18 we look at a function called s, and in Definition 19 and Theorem 20 we look at a function called S. Note that in this section we use the short-hands $\sum_\ell$ and $\sum_{\ell\neq\ell^*}$ for $\sum_{\ell\in[n]}$ and $\sum_{\ell\in[n]:\,\ell\neq\ell^*}$, respectively.

Definition 17 *Let s be the function*

$$s :\ [0,1] \to \mathbb{R}, \quad \xi \mapsto -\xi\log(\xi) + (1-\xi)\log(1-\xi).$$

Note that in contrast to the binary entropy function, there is a plus sign (not a minus sign) in front of the second term. □

Lemma 18 *The function s that is specified in Definition 17 has the following properties.*

- *As can be seen from Figure 2 (left), the graph of the function s is s-shaped.*

- *The first-order derivative of s is*

$$\frac{\mathrm{d}}{\mathrm{d}\xi} s(\xi) = -2 - \log\big(\xi(1-\xi)\big).$$

- *The second-order derivative of s is*

$$\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} s(\xi) = -\frac{1}{\xi} + \frac{1}{1-\xi} = -\frac{1-2\xi}{\xi(1-\xi)}.$$

  Clearly, the function $s(\xi)$ is strictly concave in the interval $0 \leqslant \xi < 1/2$ and strictly convex in the interval $1/2 < \xi \leqslant 1$.
- *The graph of s has a point-symmetry at $(1/2, 0)$.*

Proof: The proof of this lemma is based on straightforward calculus and is therefore omitted. □

Definition 19 *Let S be the function*

$$S: \ \Pi_{[n]} \to \mathbb{R}, \ \boldsymbol{\xi} \mapsto \sum_{\ell} s(\xi_\ell) = -\sum_{\ell} \xi_\ell \log(\xi_\ell) + \sum_{\ell} (1-\xi_\ell)\log(1-\xi_\ell).$$

□

Figure 2 (right) shows the function $S(\boldsymbol{\xi})$ for $n = 3$. More precisely, that plot shows the contour plot of the function $S(\xi_1, \xi_2, 1-\xi_1-\xi_2)$.

Clearly, if the domain of the function S were the set $[0,1]^n$, then S would not be concave everywhere because s is not concave everywhere. Therefore, the observation that is made in the following theorem, namely that S is concave, is non-trivial.[8]

Theorem 20 *The function S from Definition 19 is concave and satisfies $S(\boldsymbol{\xi}) \geqslant 0$ for all $\boldsymbol{\xi} \in \Pi_{[n]}$. Moreover,*

- *For $n = 2$, it holds that $S(\boldsymbol{\xi}) = 0$ for all $\boldsymbol{\xi} \in \Pi_{[n]}$.*
- *For $n \geqslant 3$, the function S is at almost all points in its domain a strictly concave function. However there are points in its domain and corresponding directions in which the function S is linear.*

Proof: See Appendix A. □

Lemma 21 *The Bethe entropy function can be expressed in terms of the function S as follows*

$$\begin{aligned} H_{\mathrm{B}} : \Gamma_{n \times n} &\to \mathbb{R} \\ \boldsymbol{\gamma} &\mapsto \frac{1}{2}\sum_i S(\boldsymbol{\gamma}_i) + \frac{1}{2}\sum_j S(\boldsymbol{\gamma}_j). \end{aligned}$$

[8]Because the function s is concave in $[0, 1/2]$, the function S is concave in $\Pi_{[n]} \cap [0,1/2]^n$. Therefore, as we will see, most of the work in the proof of the upcoming theorem will be devoted to proving the concavity of the function S in $\Pi_{[n]} \setminus [0,1/2]^n$.

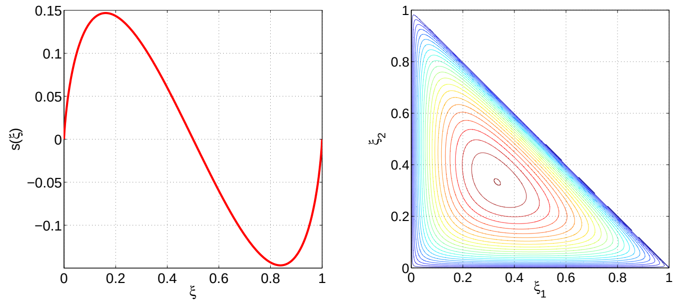


Fig. 2. Left: plot of the function s, *cf.* Definition 17. Right: contour plot of the function $S(\xi_1, \xi_2, 1-\xi_1-\xi_2)$, *cf.* Definition 19.

Proof: This result follows from

$$\begin{aligned} &H_{\mathrm{B}}(\boldsymbol{\gamma}) \\ &\overset{\text{(a)}}{=} -\sum_{i,j} \gamma_{i,j}\log(\gamma_{i,j}) + \sum_{i,j}(1-\gamma_{i,j})\log(1-\gamma_{i,j}) \\ &= \frac{1}{2}\sum_i \left(-\sum_j \gamma_{i,j}\log(\gamma_{i,j}) + \sum_j (1-\gamma_{i,j})\log(1-\gamma_{i,j}) \right) + \\ &\quad \frac{1}{2}\sum_j \left(-\sum_i \gamma_{i,j}\log(\gamma_{i,j}) + \sum_i (1-\gamma_{i,j})\log(1-\gamma_{i,j}) \right) \\ &\overset{\text{(b)}}{=} \frac{1}{2}\sum_i S(\boldsymbol{\gamma}_i) + \frac{1}{2}\sum_j S(\boldsymbol{\gamma}_j), \end{aligned}$$

where at step (a) we have used Corollary 15 and where at step (b) we have used Definition 19. □

Theorem 22 *The Bethe entropy function $H_{\mathrm{B}}(\boldsymbol{\gamma})$ is a concave function of $\boldsymbol{\gamma} \in \Gamma_{n\times n}$. Moreover, for all $\boldsymbol{\gamma} \in \Gamma_{n\times n}$ it holds that $H_{\mathrm{B}}(\boldsymbol{\gamma}) \geqslant 0$.*

Proof: Lemma 21 showed that $H_{\mathrm{B}}(\boldsymbol{\gamma})$ can be written as a sum of S-functions. The concavity of $H_{\mathrm{B}}(\boldsymbol{\gamma})$ then follows from Theorem 20 and the fact that the sum of concave functions is a concave function. Similarly, the non-negativity of $H_{\mathrm{B}}(\boldsymbol{\gamma})$ follows from Theorem 20 and the fact that the sum of non-negative functions is a non-negative function. □

Corollary 23 *The Bethe free energy function $F_{\mathrm{B}}(\boldsymbol{\gamma})$ is a convex function of $\boldsymbol{\gamma} \in \Gamma_{n\times n}$.*

Proof: This follows from $F_{\mathrm{B}}(\boldsymbol{\gamma}) = U_{\mathrm{B}}(\boldsymbol{\gamma}) - H_{\mathrm{B}}(\boldsymbol{\gamma})$ (*cf.* Corollary 15), from the fact that $U_{\mathrm{B}}(\boldsymbol{\gamma})$ is a linear function of $\boldsymbol{\gamma}$ (*cf.* Corollary 15), and from the fact that $H_{\mathrm{B}}(\boldsymbol{\gamma})$ is a concave function of $\boldsymbol{\gamma}$ (*cf.* Theorem 22). □

C. Behavior of the Bethe Entropy Function and the Bethe Free Energy Function at a Vertex of their Domain

In this section we study the Bethe entropy function and the Bethe free energy function near a vertex of their domain. Because both functions can be expressed in terms of the function S, we first study the behavior of S near a vertex of its domain.

Lemma 24 *Let*

$$\boldsymbol{\xi}(t) \triangleq \boldsymbol{\xi} + t \cdot \hat{\boldsymbol{\xi}},$$

where the vector $\boldsymbol{\xi} \in \Pi_{[n]}$ is a vertex of $\Pi_{[n]}$ and where $\hat{\boldsymbol{\xi}}$ is such that $\boldsymbol{\xi}(t) \in \Pi_{[n]}$ for small non-negative t. This means that there is an $\ell^* \in [n]$ such that $\boldsymbol{\xi}$ satisfies $\xi_{\ell^*} = 1$ and $\xi_\ell = 0$, $\ell \neq \ell^*$, and such that $\hat{\boldsymbol{\xi}}$ satisfies $\hat{\xi}_{\ell^*} < 0$, $\hat{\xi}_\ell \geqslant 0$, $\ell \neq \ell^*$, and $\sum_\ell \hat{\xi}_\ell = 0$. Then, for $0 < t \ll 1$, we have

$$S\big(\boldsymbol{\xi}(t)\big) = t \cdot |\hat{\xi}_{\ell^*}| \cdot \left(- \sum_{\ell \neq \ell'} \frac{|\hat{\xi}_\ell|}{|\hat{\xi}_{\ell^*}|} \log \left(\frac{|\hat{\xi}_\ell|}{|\hat{\xi}_{\ell^*}|} \right) \right) + O(t^2), \tag{4}$$

i.e., *the function $S\big(\boldsymbol{\xi}(t)\big)$ can very well be approximated by a linear function for $0 < t \ll 1$. Note that the coefficient of t in* (4) *is non-negative.*

Proof: See Appendix B. □

A word of caution: the behavior of the function S is somewhat special around a vertex $\boldsymbol{\xi}$ of $\Pi_{[n]}$: namely, *in general there is no gradient vector $\boldsymbol{G}$* such that $S(\boldsymbol{\xi} + t \cdot \hat{\boldsymbol{\xi}}) = S(\boldsymbol{\xi}) + t \cdot \sum_\ell G_\ell \hat{\xi}_\ell + O(t^2) = t \cdot \sum_\ell G_\ell \hat{\xi}_\ell + O(t^2)$ for $0 < t \ll 1$ and for all possible direction vectors $\hat{\boldsymbol{\xi}}$.

Lemma 24 has the following consequences for the behavior of the Bethe entropy function at a vertex of its domain.

Lemma 25 *Let*

$$\boldsymbol{\gamma}(t) \triangleq \boldsymbol{\gamma} + t \cdot \hat{\boldsymbol{\gamma}},$$

where $\boldsymbol{\gamma} \in \mathcal{C}_{\mathcal{E}}$ is a vertex of $\Gamma_{n \times n}$ and where $\hat{\boldsymbol{\gamma}}$ is such that $\boldsymbol{\gamma}(t) \in \Gamma_{n \times n}$ for small non-negative t. This means that $\boldsymbol{\gamma}$ corresponds to the permutation $\sigma_{\boldsymbol{\gamma}}$. (In the following statement we will use the short-hands $\sigma \triangleq \sigma_{\boldsymbol{\gamma}}$ and $\bar{\sigma} \triangleq \sigma_{\boldsymbol{\gamma}}^{-1}$.) Then, for $0 < t \ll 1$, we have

$$\begin{aligned} &H_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big) \\ &= t \sum_i |\hat{\gamma}_{i,\sigma(i)}| \cdot \left(- \sum_{j \neq \sigma(i)} \frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{i,\sigma(i)}|} \log \left(\frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{i,\sigma(i)}|} \right) \right) + O(t^2) \\ &= t \sum_j |\hat{\gamma}_{\bar{\sigma}(j),j}| \cdot \left(- \sum_{i \neq \bar{\sigma}(j)} \frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{\bar{\sigma}(j),j}|} \log \left(\frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{\bar{\sigma}(j),j}|} \right) \right) + O(t^2), \end{aligned}$$

i.e., *the function $H_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big)$ can very well be approximated by a linear function for $0 < t \ll 1$. Note that the coefficient of t is non-negative.*

Proof: See Appendix C. □

Assume that $\hat{\boldsymbol{\gamma}}$ in Lemma 25 is chosen such that $\sum_i |\hat{\gamma}_{i,\sigma(i)}| = 1$. (If this is not the case, then $\hat{\boldsymbol{\gamma}}$ can be rescaled by a positive real number such that this condition is satisfied.) The coefficient of t in the first display equation of Lemma 25 can be given the following meaning. It is the entropy rate of the time-invariant Markov chain corresponding to the (back-trackless) random walk on the NFG $\mathsf{N}(\boldsymbol{\theta})$ (*cf.* Figure 1) with the following properties:[9]

- The probability of being at vertex $i \in \mathcal{I}$ is $|\hat{\gamma}_{i,\sigma(i)}|$.
- The probability of going to vertex $j \in \mathcal{J} \setminus \{\sigma(i)\}$, conditioned on being at vertex $i \in \mathcal{I}$, is $|\hat{\gamma}_{i,j}|/|\hat{\gamma}_{i,\sigma(i)}|$.
  The probability of going to vertex $\sigma(i) \in \mathcal{J}$, conditioned on being at vertex $i \in \mathcal{I}$, is 0.
- The probability of being at vertex $j \in \mathcal{J}$ is $|\hat{\gamma}_{\bar{\sigma}(j),j}|$.
- The probability of going to vertex $\bar{\sigma}(j) \in \mathcal{I}$, conditioned on being at vertex $j \in \mathcal{J}$, is 1.
  The probability of going to vertex $i' \in \mathcal{I} \setminus \{\bar{\sigma}(j)\}$, conditioned on being at vertex $j \in \mathcal{J}$, is 0.

The above two half-steps of the random walk can be combined into one step.

- The probability of being at vertex $i \in \mathcal{I}$ is $|\hat{\gamma}_{i,\sigma(i)}|$.
- For $i, i' \in \mathcal{I}$ with $i \neq i'$, the probability of going to vertex $\sigma(i')$ and then to vertex i', conditioned on being at vertex i, is $|\hat{\gamma}_{i,\sigma(i')}|/|\hat{\gamma}_{i,\sigma(i)}|$.

An analogous interpretation can be given to the coefficient of t in the second display equation of Lemma 25. Observe that the condition $\sum_i |\hat{\gamma}_{i,\sigma(i)}| = 1$ is equivalent to the condition $\sum_j |\hat{\gamma}_{\bar{\sigma}(j),j}| = 1$.

Note that similar random walks appeared in the analysis of the Bethe entropy function for so-called cycle codes (*cf.* [32]) and in the analysis of linear programming decoding of low-density parity-check codes (*cf.* [33], which gives a random walk interpretation of a result by Arora, Daskalakis, Steurer [34] and its extensions by Halabi and Even [35]). Actually, given the fact that the symmetric difference of two perfect matchings corresponds to a union of cycles in $\mathsf{N}(\boldsymbol{\theta})$, the similarity of the random walks here and of the random walks in the above-mentioned context of cycle codes is not totally surprising.

We come now to the main result of this subsection. Although this result is interesting in its own right, it will be especially important for the convergence analysis of the SPA in Section V.

Theorem 26 *Let*

$$\boldsymbol{\gamma}(t) \triangleq \boldsymbol{\gamma} + t \cdot \hat{\boldsymbol{\gamma}},$$

where $\boldsymbol{\gamma} \in \mathcal{C}_{\mathcal{E}}$ is a vertex of $\Gamma_{n \times n}$ and where $\hat{\boldsymbol{\gamma}}$ is such that $\boldsymbol{\gamma}(t) \in \Gamma_{n \times n}$ for small non-negative t. This means that $\boldsymbol{\gamma}$ corresponds to the permutation $\sigma_{\boldsymbol{\gamma}}$. (In the following statement we will use the short-hands $\sigma \triangleq \sigma_{\boldsymbol{\gamma}}$ and $\bar{\sigma} \triangleq \sigma_{\boldsymbol{\gamma}}^{-1}$.) We also assume that $\hat{\boldsymbol{\gamma}}$ is normalized as follows

$$\sum_i |\hat{\gamma}_{i,\sigma(i)}| = \sum_j |\hat{\gamma}_{\bar{\sigma}(j),j}| = 1. \tag{5}$$

Then, for $0 < t \ll 1$, we have

$$F_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big) \geqslant - \sum_i \log(\theta_{i,\sigma(i)}) - t \cdot \log(\rho) + O(t^2), \tag{6}$$

where ρ is the maximal (real) eigenvalue of the $n \times n$ matrix $\boldsymbol{A}$ with entries

$$A_{i,i'} \triangleq \begin{cases} \frac{\theta_{i,\sigma(i')}}{\theta_{i,\sigma(i)}} & \textit{(if } i \neq i'\textit{)} \\ 0 & \textit{(otherwise)} \end{cases}.$$

[9] For a discussion of the entropy rate of a time-invariant Markov chain, see, *e.g.*, [31, Section 4.2].

Note that equality holds in (6) *for the matrix* $\hat{\boldsymbol{\gamma}}$ *with entries*

$$\hat{\gamma}_{i,\sigma(i')} \triangleq \begin{cases} +\kappa \cdot \frac{u_i^{\mathrm{L}} \cdot A_{i,i'} \cdot u_{i'}^{\mathrm{R}}}{\rho} & \text{(if } i \neq i'\text{)} \\ -\kappa \cdot u_i^{\mathrm{L}} \cdot u_i^{\mathrm{R}} & \text{(otherwise)} \end{cases},$$

where $\boldsymbol{u}^{\mathrm{L}}$ *and* $\boldsymbol{u}^{\mathrm{R}}$ *are, respectively, the left and right eigenvectors of* $\boldsymbol{A}$ *with eigenvalue* ρ*, and where* κ *is a suitable normalization constant such that* (5) *is satisfied.*

Proof: See Appendix D. □

Corollary 27 *Consider a vertex* $\boldsymbol{\gamma}$ *of* $\Gamma_{n\times n}$ *and define* ρ *for* $\boldsymbol{\gamma}$ *as in Theorem 26.*

- *If* $\rho < 1$ *then* F_{B} *has its unique minimum at* $\boldsymbol{\gamma}$*.*
- *If* $\rho > 1$ *then* F_{B} *is not minimal at* $\boldsymbol{\gamma}$*.*

Proof: From Theorem 26 we know that

$$F_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big) \geqslant -\sum_i \log(\theta_{i,\sigma(i)}) - t \cdot \log(\rho) + O(t^2),$$

with equality for the direction matrix $\hat{\boldsymbol{\gamma}}$ that was specified there. Moreover, from Corollary 23 we know that F_{B} is convex over $\Gamma_{n\times n}$. Therefore, if $\log(\rho) < 0$ (*i.e.*, $\rho < 1$) then F_{B} has a unique minimum at $\boldsymbol{\gamma}$. On the other hand, if $\log(\rho) > 0$ (*i.e.*, $\rho > 1$) then F_{B} cannot be minimal at $\boldsymbol{\gamma}$.

Note that for $\log(\rho) = 0$ (*i.e.*, $\rho = 1$), the minimality / non-minimality of F_{B} at $\boldsymbol{\gamma}$ is determined by the $O(t^2)$ term. □

Typically, the Bethe entropy function and the Bethe free energy function have positive or negative infinite slope at a vertex of their domain because of the appearance of terms like $c \cdot t \cdot \log(t)$. However, because for the function S all these $c \cdot t \cdot \log(t)$ terms cancel in the vicinity of a vertex of its domain (see the proof of Theorem 20, in particular Eq. (17) in Appendix A-B), the slopes of the Bethe entropy function and the Bethe free energy function are finite at a vertex of their domain.

Let us conclude this section by pointing out that the observations that were made in this subsection give an alternative viewpoint of some of the results that were presented in [14, Section 3].

V. Sum-Product-Algorithm-Based Search of the Minimum of the Bethe Free Energy Function

Assumption 28 *In this section we make the following two assumptions, both with the goal of simplifying the wording of most results without hurting their generality too much.*

- *We assume that* $n \geqslant 2$ *and that* $\boldsymbol{\theta}$ *is a* positive *matrix of size* $n \times n$*. In that respect, see also the comments in Assumption 16.*
- *We assume that the minimum of the Bethe free energy function* F_{B} *is either in the interior of* $\Gamma_{n\times n}$ *or at a vertex of* $\Gamma_{n\times n}$*, but not at a non-vertex boundary point of* $\Gamma_{n\times n}$*. A possibility to guarantee this with probability* 1 *is to apply tiny random perturbations to the entries of* $\boldsymbol{\theta}$*.* □

In Definition 12 we have defined the Bethe permanent of a square matrix $\boldsymbol{\theta}$ via the minimum of the Bethe free energy function of the NFG $\mathsf{N}(\boldsymbol{\theta})$. In Corollary 23 we have seen that the Bethe free energy function is a convex function, *i.e.*, it behaves very favorably. This means that we could use any generic optimization algorithm (see, *e.g.*, [26], [36]) to find the minimum of the Bethe free energy function, and with that the Bethe permanent of $\boldsymbol{\theta}$. However, given the special structure of the optimization problem, there is the hope that there are more efficient approaches.

A natural candidate for searching this minimum is the SPA [28]–[30]. The reason for this is that a theorem by Yedidia, Freeman, and Weiss [9] says that fixed points of the SPA correspond to stationary points of the Bethe free energy function.[10] Given the convexity of the Bethe free energy function, the following two questions must therefore be answered:

- If the minimum of F_{B} is in the interior of $\Gamma_{n\times n}$, does the SPA always converge to a fixed point?
- If the minimum of F_{B} is at a vertex of $\Gamma_{n\times n}$, does the SPA find that vertex?

In this section we answer both questions affirmatively, independently of the matrix $\boldsymbol{\theta}$, and (nearly) independently of the chosen initial messages.

The rest of this section is structured as follows. First we discuss the details of the SPA message update rules in Section V-A. Afterwards, we state the SPA convergence result in Section V-B.

A. Sum-Product Algorithm Message Update Rules

In this subsection we derive the SPA message update rules for the NFG $\mathsf{N}(\boldsymbol{\theta})$ in Figure 1. Here we only give the technical details; for a general discussion w.r.t. motivations behind the SPA we refer to [28]–[30]. Note that in contrast to [8] we use an undampened version of the SPA.

On a high level, the SPA works as follows. With every edge in Figure 1 we associate a right-going message and a left-going message. Every iteration of the SPA consists then of two half-iterations, in the first half-iteration the right-going messages are updated based on the left-going messages and in the second half-iteration the left-going messages are updated based on the right-going messages. Finally, once some suitable convergence criterion is met or a fixed number of iterations has been reached, the pseudo-marginal vector (belief vector) is computed based on the messages at the last iteration.

Mathematically, we define for every $t \geqslant 0$ and every edge $(i,j) \in \mathcal{I} \times \mathcal{J}$ a left-going message $\overleftarrow{\mu}_{i,j}^{(t)} : \mathcal{A}_{i,j} \to \mathbb{R}$, and for every $t \geqslant 1$ and every edge $(i,j) \in \mathcal{I} \times \mathcal{J}$ a right-going message $\overrightarrow{\mu}_{i,j}^{(t)} : \mathcal{A}_{i,j} \to \mathbb{R}$.

For every left-going and for every right-going message it

[10] Strictly speaking, for NFGs with hard constraints, *i.e.*, NFGs that contain local functions that can assume the value zero for certain points in their domain (which is the case for $\mathsf{N}(\boldsymbol{\theta})$), this statement has only been proven for *interior* stationary points of the Bethe free energy (*cf.* [9, Theorem 2]). For SPA fixed points with some beliefs equal to zero it is only conjectured that they correspond to edge-stationary points of the Bethe free energy function (*cf.* discussion in [9, Section VI.D]).

turns out to be sufficient to keep track of the likelihood ratios

$$\overrightarrow{\Lambda}_{i,j}^{(t)} \triangleq \frac{\overrightarrow{\mu}_{i,j}^{(t)}(0)}{\overrightarrow{\mu}_{i,j}^{(t)}(1)}, \qquad \overleftarrow{\Lambda}_{i,j}^{(t)} \triangleq \frac{\overleftarrow{\mu}_{i,j}^{(t)}(0)}{\overleftarrow{\mu}_{i,j}^{(t)}(1)},$$

respectively. Actually, for the NFG under consideration it is more convenient to deal with the inverses of these quantities, and so we define the inverse likelihood ratios as follows

$$\overrightarrow{\mathrm{V}}_{i,j}^{(t)} \triangleq \left(\overrightarrow{\Lambda}_{i,j}^{(t)}\right)^{-1}, \qquad \overleftarrow{\mathrm{V}}_{i,j}^{(t)} \triangleq \left(\overleftarrow{\Lambda}_{i,j}^{(t)}\right)^{-1}.$$

Lemma 29 *Consider the NFG $\mathsf{N}(\boldsymbol{\theta})$. The inverse likelihood ratio update rules for the left-hand side and right-hand side function nodes of $\mathsf{N}(\boldsymbol{\theta})$ are given by, respectively,*

$$\overrightarrow{\mathrm{V}}_{i,j}^{(t)} = \frac{\sqrt{\theta_{i,j}}}{\sum_{j' \neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mathrm{V}}_{i,j'}^{(t-1)}}, \quad t \geqslant 1,\ (i,j) \in \mathcal{I} \times \mathcal{J},$$

$$\overleftarrow{\mathrm{V}}_{i,j}^{(t)} = \frac{\sqrt{\theta_{i,j}}}{\sum_{i' \neq i} \sqrt{\theta_{i',j}} \cdot \overrightarrow{\mathrm{V}}_{i',j}^{(t)}}, \qquad t \geqslant 1,\ (i,j) \in \mathcal{I} \times \mathcal{J}.$$

The pseudo-marginal vector at the left-hand side and right-hand side function nodes of $\mathsf{N}(\boldsymbol{\theta})$ are given by, respectively,

$$\beta_{i,\boldsymbol{a}_i}^{(t)}\Big|_{\boldsymbol{a}_i = \boldsymbol{u}_j} \propto \sqrt{\theta_{i,j}} \cdot \overleftarrow{\mathrm{V}}_{i,j}^{(t)}, \quad t \geqslant 0,\ (i,j) \in \mathcal{I} \times \mathcal{J},$$

$$\beta_{j,\boldsymbol{a}_j}^{(t)}\Big|_{\boldsymbol{a}_j = \boldsymbol{u}_i} \propto \sqrt{\theta_{i,j}} \cdot \overrightarrow{\mathrm{V}}_{i,j}^{(t)}, \quad t \geqslant 1,\ (i,j) \in \mathcal{I} \times \mathcal{J}.$$

Here the proportionality constants are defined such that for every function node the beliefs sum to 1.

Proof: See Appendix E. □

Let us remark on the side that the above update equations can be reformulated such that we only multiply by factors like $\theta_{i,j}$ instead of by factors like $\sqrt{\theta_{i,j}}$. We leave the details to the reader.

Remark 30 *The SPA messages for the NFG $\mathsf{N}(\boldsymbol{\theta})$ exhibit the following property, a property that we will henceforth call "message gauge invariance." Namely, consider the messages*

$$\left\{\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\right\}_{i,j,t} \quad \text{and} \quad \left\{\overrightarrow{\mathrm{V}}_{i,j}^{(t)}\right\}_{i,j,t}$$

that are connected by the update equations in Lemma 29. It is then easy to show that for any $C \in \mathbb{R}_{>0}$ the messages

$$\left\{C \cdot \overleftarrow{\mathrm{V}}_{i,j}^{(t)}\right\}_{i,j,t} \quad \text{and} \quad \left\{\frac{1}{C} \cdot \overrightarrow{\mathrm{V}}_{i,j}^{(t)}\right\}_{i,j,t}$$

also satisfy the update equations in Lemma 29. Moreover, the pseudo-marginals $\{\beta_{i,\boldsymbol{a}_i}^{(t)}\}_{i,\boldsymbol{a}_i,t}$ and $\{\beta_j^{(t)}(\boldsymbol{a}_j)\}_{j,\boldsymbol{a}_j,t}$ are left unchanged by this rescaling of the inverse likelihood ratios. This is because the normalization that appears in the definition of $\{\beta_{i,\boldsymbol{a}_i}^{(t)}\}_{i,\boldsymbol{a}_i,t}$ and $\{\beta_j^{(t)}(\boldsymbol{a}_j)\}_{j,\boldsymbol{a}_j,t}$ removes the influence of this message rescaling. □

Strictly speaking, the Bethe free energy function can only be evaluated at fixed points of the SPA. However, very often it is desirable to track the progress towards the minimal Bethe free energy function value. This can be done via the so-called pseudo-dual function of the Bethe free energy function [37], [38]. This function has the following two properties: it can be evaluated at any point during the SPA computations, and at a fixed point of the SPA its value equals the value of the Bethe free energy function. However, in general it is *not* a non-increasing or a non-decreasing function of the iteration number.

Lemma 31 *Consider the NFG $\mathsf{N}(\boldsymbol{\theta})$. For any set of left-going messages $\{\overleftarrow{\mathrm{V}}_{i,j}\}_{i,j}$ and any set of right-going messages $\{\overrightarrow{\mathrm{V}}_{i,j}\}_{i,j}$, the pseudo-dual function of the Bethe free energy function is*

$$\begin{aligned} F_{\mathrm{Bethe}}^{\#}\left(\{\overleftarrow{\mathrm{V}}_{i,j}\}, \{\overrightarrow{\mathrm{V}}_{i,j}\}\right) = &-\sum_i \log\left(\sum_j \sqrt{\theta_{i,j}} \cdot \overleftarrow{\mathrm{V}}_{i,j}\right) \\ &-\sum_j \log\left(\sum_i \sqrt{\theta_{i,j}} \cdot \overrightarrow{\mathrm{V}}_{i,j}\right) \\ &+\sum_{i,j} \log\left(1 + \overleftarrow{\mathrm{V}}_{i,j} \cdot \overrightarrow{\mathrm{V}}_{i,j}\right) \end{aligned}$$

Proof: See Appendix F. □

In particular, if desired, we can evaluate $F_{\mathrm{Bethe}}^{\#}$ after every half-iteration of the SPA, *i.e.*, we can compute $F_{\mathrm{Bethe}}^{\#}(\{\overleftarrow{\mathrm{V}}_{i,j}^{(t-1)}\}, \{\overrightarrow{\mathrm{V}}_{i,j}^{(t)}\})$ and $F_{\mathrm{Bethe}}^{\#}(\{\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\}, \{\overrightarrow{\mathrm{V}}_{i,j}^{(t)}\})$ for every $t \geqslant 1$.

B. Convergence of the Sum-Product Algorithm

Note that there are rather few general results concerning the behavior of message-passing type algorithms for NFGs with cycles. For certain classes of graphical models and message-passing type algorithms, early results showed that under the assumption that the algorithm converges then the obtained estimates are correct (see, *e.g.*, the results in [39], [40]). Later, conditions for convergence were established for a variety of graphical models and message-passing type algorithms (see, *e.g.*, [41]–[44] and references therein). However, these results do not seem to be applicable to the NFG under consideration in this paper.

The SPA convergence proof that is the most relevant for the present paper is the one in the paper by Bayati and Nair [18] (see also the comments that we made about this paper in Section I-C). However, the fact that the graphical model in [18] counts matchings (and not only perfect matchings like here), implies a different behavior of the Bethe free energy function near the boundary of its domain, and so no separate analysis of interior and boundary minima of the Bethe free energy is required in the convergence proof in [18].

Note that, interestingly enough, establishing convergence for the SPA on $\mathsf{N}(\boldsymbol{\theta})$ is independent of the choice of $\boldsymbol{\theta}$, which is in contrast to, say, Gaussian graphical models where the convergence behavior not only depends on the connectivity of the underlying graph but also on the values of the non-zero entries of the information matrix describing the Gaussian graphical model. (Of course, the convergence *speed* of the SPA on $\mathsf{N}(\boldsymbol{\theta})$ does depend on the choice of $\boldsymbol{\theta}$.)

Theorem 32 *Consider the SPA for NFG $\mathsf{N}(\boldsymbol{\theta})$, for which the message update rules were established in Lemma 29. For any initial set of inverse likelihood ratios $\big\{\overleftarrow{\mathrm{V}}_{i,j}^{(0)}\big\}_{i,j}$ that satisfies $0 < \overleftarrow{\mathrm{V}}_{i,j}^{(0)} < \infty$, $(i,j) \in \mathcal{I} \times \mathcal{J}$, the pseudo-marginals computed by the SPA converge to the pseudo-marginals that minimize the Bethe free energy function of $\mathsf{N}(\boldsymbol{\theta})$. More precisely, we can make the following statements.*[11]

- *If the minimum of F_{B} is in the interior of $\Gamma_{n\times n}$, then the inverse likelihood ratios*
$$\big\{\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\big\}_{i,j,t}\Big|_{t\to\infty} \text{ and } \big\{\overrightarrow{\mathrm{V}}_{i,j}^{(t)}\big\}_{i,j,t}\Big|_{t\to\infty}$$
stay bounded and converge (modulo the message gauge invariance mentioned in Remark 30) to the fixed point inverse likelihood ratios corresponding to the minimum of F_{B}.
- *If the minimum of F_{B} is at at the vertex $\boldsymbol{\gamma}$ of $\Gamma_{n\times n}$, then the inverse likelihood ratios satisfy*
$$\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\Big|_{j=\sigma_{\boldsymbol{\gamma}}(i)} \xrightarrow{t\to\infty} \infty, \qquad \overrightarrow{\mathrm{V}}_{i,j}^{(t)}\Big|_{j=\sigma_{\boldsymbol{\gamma}}(i)} \xrightarrow{t\to\infty} \infty,$$
$$\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\Big|_{j\neq\sigma_{\boldsymbol{\gamma}}(i)} \xrightarrow{t\to\infty} 0, \qquad \overrightarrow{\mathrm{V}}_{i,j}^{(t)}\Big|_{j\neq\sigma_{\boldsymbol{\gamma}}(i)} \xrightarrow{t\to\infty} 0.$$

Finally,
$$\left|\exp\left(-F_{\mathrm{Bethe}}^{\#}\Big(\big\{\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\big\},\big\{\overrightarrow{\mathrm{V}}_{i,j}^{(t)}\big\}\Big)\right) - \mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})\right| \leqslant C \cdot \mathrm{e}^{-\nu\cdot t}$$
for some constants $C, \nu \in \mathbb{R}_{>0}$ that depend on the matrix $\boldsymbol{\theta}$ and the initial messages.

Proof: See Appendix G. □

Explicit convergence speed estimates (in particular, values for C and ν) can be extracted from the proof of Theorem 32. However, we think that a more sophisticated analysis might yield tighter convergence speed estimates; we leave this as an open problem for future research.

VI. Finite-Graph-Cover Interpretation of the Bethe Permanent

Note that the definition of the permanent of $\boldsymbol{\theta}$ in Definition 1 has a "combinatorial flavor." In particular, it can be seen as a sum over all weighted perfect matchings of a complete bipartite graph. This is in contrast to the definition of the Bethe permanent of $\boldsymbol{\theta}$ (*cf.* Definitions 11 and 12) that has an "analytical flavor." In this section we show that it is possible to represent the Bethe permanent by an expression that has a "combinatorial flavor." We do this by applying the results from [11], that hold for general NFGs, to the NFG $\mathsf{N}(\boldsymbol{\theta})$. The key concept in that respect are so-called finite graph covers. (We keep the discussion here somewhat brief and we refer to [11] for all the details. See also [45].)

Definition 33 (see, *e.g.*, [46], [47]) *A* cover *of a graph G with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$ is a graph $\widetilde{\mathsf{G}}$ with vertex set $\widetilde{\mathcal{V}}$ and edge set $\widetilde{\mathcal{E}}$, along with a surjection $\pi : \widetilde{\mathcal{V}} \to \mathcal{V}$ which is a graph homomorphism (*i.e., *π takes adjacent vertices of $\widetilde{\mathsf{G}}$ to adjacent vertices of G) such that for each vertex $v \in \mathcal{V}$ and each $\widetilde{v} \in \pi^{-1}(v)$, the neighborhood $\partial(\widetilde{v})$ of $\widetilde{v}$ is mapped bijectively to $\partial(v)$. A cover is called an M-*cover*, where $M \in \mathbb{Z}_{>0}$, if $\left|\pi^{-1}(v)\right| = M$ for every vertex v in $\mathcal{V}$.*[12] □

[11] We remind the reader of the assumptions that were made in Assumption 28.

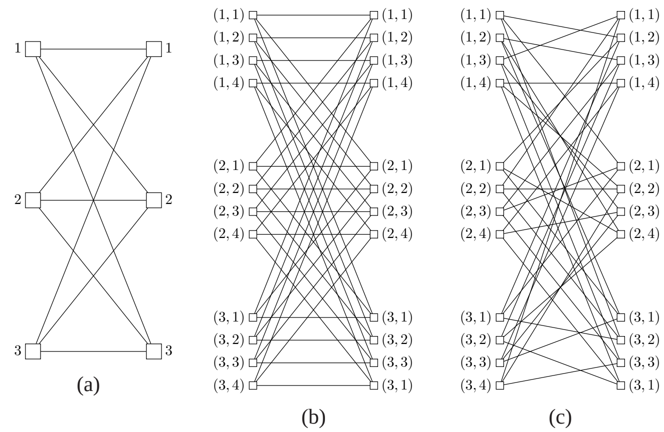


Fig. 3. (a) NFG $\mathsf{N}(\boldsymbol{\theta})$ for $n = 3$. (b) "Trivial" 4-cover of $\mathsf{N}(\boldsymbol{\theta})$ (c) A possible 4-cover of $\mathsf{N}(\boldsymbol{\theta})$.

Because NFGs are graphs, it is straightforward to extend this definition to NFGs. (Of course, the variables that are associated with the M copies of an edge are allowed to take on different values.) For an M-cover, the left-hand side function nodes will be labeled by elements of $\mathcal{I} \times [M]$, the right-hand side function nodes will be labeled by elements of $\mathcal{J} \times [M]$, and the edges will be labeled by elements of a cover-dependent subset of $\mathcal{I} \times [M] \times \mathcal{J} \times [M]$.

Example 34 *Let $n = 3$. The NFG $\mathsf{N}(\boldsymbol{\theta})$ is shown in Figure 3(a). There is only one 1-cover of $\mathsf{N}(\boldsymbol{\theta})$, namely $\mathsf{N}(\boldsymbol{\theta})$ itself. Two possible 4-covers of $\mathsf{N}(\boldsymbol{\theta})$ are shown in Figures 3(b)–(c). The 4-cover in Figure 3(b) is "trivial" in the sense that it consists of 4 disconnected copies of $\mathsf{N}(\boldsymbol{\theta})$. On the other hand, the 4-cover in Figure 3(c) is "nontrivial" in the sense that it consists of 4 copies of $\mathsf{N}(\boldsymbol{\theta})$ that are intertwined.* □

Lemma 35 *Let $\widetilde{\mathcal{N}}_M(\boldsymbol{\theta})$ be the set of all M-covers $\widetilde{\mathsf{N}}$ of $\mathsf{N}(\boldsymbol{\theta})$. It holds that*
$$\left|\widetilde{\mathcal{N}}_M(\boldsymbol{\theta})\right| = (M!)^{(n^2)}. \tag{7}$$

Proof: This follows from [11, Lemma 15] and the fact that the NFG $\mathsf{N}(\boldsymbol{\theta})$ has n^2 full edges. □

The following definition is the main definition of this section.

Definition 36 *For any $M \in \mathbb{Z}_{>0}$ we define the degree-M Bethe permanent of $\boldsymbol{\theta}$ to be*
$$\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \triangleq \sqrt[M]{\Big\langle Z_{\mathrm{G}}(\widetilde{\mathsf{N}})\Big\rangle_{\widetilde{\mathsf{N}}\in\widetilde{\mathcal{N}}_M}},$$

[12] The number M is also known as the degree of the cover. (Not to be confused with the degree of a vertex.)

where the angular brackets represent the arithmetic average of $Z_{\mathrm{G}}(\widetilde{\mathsf{N}})$ over all $\widetilde{\mathsf{N}} \in \widetilde{\mathcal{N}}_M$. (Note that the right-hand side is based on the Gibbs partition function, not the Bethe partition function.) □

As we will now show, one can express $Z_{\mathrm{G}}(\widetilde{\mathsf{N}})$ for any M-cover $\widetilde{\mathsf{N}}$ of $\mathsf{N}(\boldsymbol{\theta})$ as the permanent of some matrix that is derived from $\boldsymbol{\theta}$.

Definition 37 *For any $M \in \mathbb{Z}_{>0}$ we define $\widetilde{\Psi}_M$ to be the set*

$$\widetilde{\Psi}_M \triangleq \left\{ \widetilde{\boldsymbol{P}} = \left\{ \widetilde{\boldsymbol{P}}^{(i,j)} \right\}_{i \in \mathcal{I}, j \in \mathcal{J}} \;\middle|\; \widetilde{\boldsymbol{P}}^{(i,j)} \in \mathcal{P}_{M \times M} \right\}.$$

Moreover, for $\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M$ we define the $\widetilde{\boldsymbol{P}}$-lifting of $\boldsymbol{\theta}$ to be the following $(nM) \times (nM)$ matrix

$$\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}} \triangleq \begin{pmatrix} \theta_{1,1} \widetilde{\boldsymbol{P}}^{(1,1)} & \cdots & \theta_{1,n} \widetilde{\boldsymbol{P}}^{(1,n)} \\ \vdots & & \vdots \\ \theta_{n,1} \widetilde{\boldsymbol{P}}^{(n,1)} & \cdots & \theta_{n,n} \widetilde{\boldsymbol{P}}^{(n,n)} \end{pmatrix}.$$

□

For any positive integer M it is straightforward to see that there is a bijection between the set $\widetilde{\mathcal{N}}_M(\boldsymbol{\theta})$ of all M-covers of $\mathsf{N}(\boldsymbol{\theta})$ and the set $\{\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\}_{\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M}$. In particular, because of Lemma 8, for an M-cover $\widetilde{\mathsf{N}}$ and its corresponding matrix $\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}$ it holds that $Z_{\mathrm{G}}(\widetilde{\mathsf{N}}) = \mathrm{perm}(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}})$. Therefore, we have the following reformulation of Definition 36.

Definition 38 (Reformulation of Definition 36) *For any $M \in \mathbb{Z}_{>0}$ we define the degree-M Bethe permanent of $\boldsymbol{\theta}$ to be*

$$\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \triangleq \sqrt[M]{\left\langle \mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}} \right) \right\rangle_{\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M}}, \tag{8}$$

where the angular brackets represent the arithmetic average of $\mathrm{perm}\big(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\big)$ over all $\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M$. (Note that the permanent, not the Bethe permanent, appears on the right-hand side of the above expression.) □

In order to better appreciate the right-hand side of the above expression, it is worthwhile to make the following two observations.

- For $M = 1$, the averaging is trivial because $\widetilde{\Psi}_M$ contains only one element. Moreover, letting $\widetilde{\boldsymbol{P}}$ be this single element, it holds that $\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}} = \boldsymbol{\theta}$. Therefore
$$\mathrm{perm}_{\mathrm{B},1}(\boldsymbol{\theta}) = \mathrm{perm}(\boldsymbol{\theta}).$$
- For any $M \in \mathbb{Z}_{>0}$, the "trivial" M-cover of $\mathsf{N}(\boldsymbol{\theta})$ is given by the choice $\widetilde{\boldsymbol{P}} = \left\{ \widetilde{\boldsymbol{P}}^{(i,j)} \right\}_{i \in \mathcal{I}, j \in \mathcal{J}}$ with $\widetilde{\boldsymbol{P}}^{(i,j)} = \widetilde{\boldsymbol{I}}$, $(i,j) \in \mathcal{I} \times \mathcal{J}$, where $\widetilde{\boldsymbol{I}}$ is the identity matrix of size $M \times M$. For this M-cover we obtain
$$\mathrm{perm}(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}) = \mathrm{perm}(\boldsymbol{\theta})^M,$$
i.e.
$$\sqrt[M]{\mathrm{perm}(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}})} = \mathrm{perm}(\boldsymbol{\theta}).$$

$$\begin{array}{ll} \mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta})\Big|_{M \to \infty} & = \mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) \\ \qquad\Big| & \\ \mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) & \\ \qquad\Big| & \\ \mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta})\Big|_{M=1} & = \mathrm{perm}(\boldsymbol{\theta}) \end{array}$$

Fig. 4. The degree-M Bethe permanent of the non-negative matrix $\boldsymbol{\theta}$ for different values of M.

With this, we are ready for the main result of this section.

Theorem 39 *It holds that*

$$\limsup_{M \to \infty} \; \mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) = \mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}).$$

Proof: This follows from Definitions 12 and 38, along with the application of [11, Theorem 19] to $\mathsf{N} = \mathsf{N}(\boldsymbol{\theta})$. □

Theorem 39, together with the relation $\mathrm{perm}_{\mathrm{B},1}(\boldsymbol{\theta}) = \mathrm{perm}(\boldsymbol{\theta})$, are visualized in Figure 4. Because the permanents that appear on the right-hand side of (8) are combinatorial objects, Definition 38 and Theorem 39 give the promised "combinatorial characterization" of the Bethe permanent.

A. The Bethe Permanent for Matrices of Size 2×2

In this and the following subsections we illustrate the concepts and results that have been presented so far in this section by having a detailed look at the case $n = 2$, *i.e.*, we study the permanent, the Bethe permanent, and the degree-M Bethe permanent for the matrix

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{1,1} & \theta_{1,2} \\ \theta_{2,1} & \theta_{2,2} \end{pmatrix}.$$

The corresponding NFG $\mathsf{N}(\boldsymbol{\theta})$ is shown in Figure 5(a). Of course, nobody would use the Bethe permanent to approximate the permanent of a 2×2 matrix, however, it gives some good insights on the strengths and the weaknesses of the Bethe approximation to the permanent.

Lemma 40 *For $n = 2$ it holds that*

$$\begin{aligned} \mathrm{perm}(\boldsymbol{\theta}) &= \theta_{1,1}\theta_{2,2} + \theta_{2,1}\theta_{1,2}, \\ \mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) &= \max(\theta_{1,1}\theta_{2,2},\ \theta_{2,1}\theta_{1,2}). \end{aligned}$$

Proof: The result for $\mathrm{perm}(\boldsymbol{\theta})$ follows from Definition 1. On the other hand, in order to obtain $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})$, we apply Corollary 15. The crucial step in Corollary 15 is to minimize $F_{\mathrm{B}}(\boldsymbol{\gamma})$ over $\boldsymbol{\gamma} \in \Gamma_{2\times 2}$. Because $H_{\mathrm{B}}(\boldsymbol{\gamma}) = 0$, $\boldsymbol{\gamma} \in \Gamma_{2\times 2}$, minimizing $F_{\mathrm{B}}(\boldsymbol{\gamma})$ is equivalent to minimizing $U_{\mathrm{B}}(\boldsymbol{\gamma}) = -\sum_{i,j} \gamma_{i,j} \log(\theta_{i,j})$.

- For $\theta_{1,1}\theta_{2,2} = \theta_{1,2}\theta_{2,1}$ the minimum is achieved at every $\boldsymbol{\gamma} \in \Gamma_{2\times 2}$.
- For $\theta_{1,1}\theta_{2,2} > \theta_{1,2}\theta_{2,1}$ the minimum is achieved at $\boldsymbol{\gamma} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

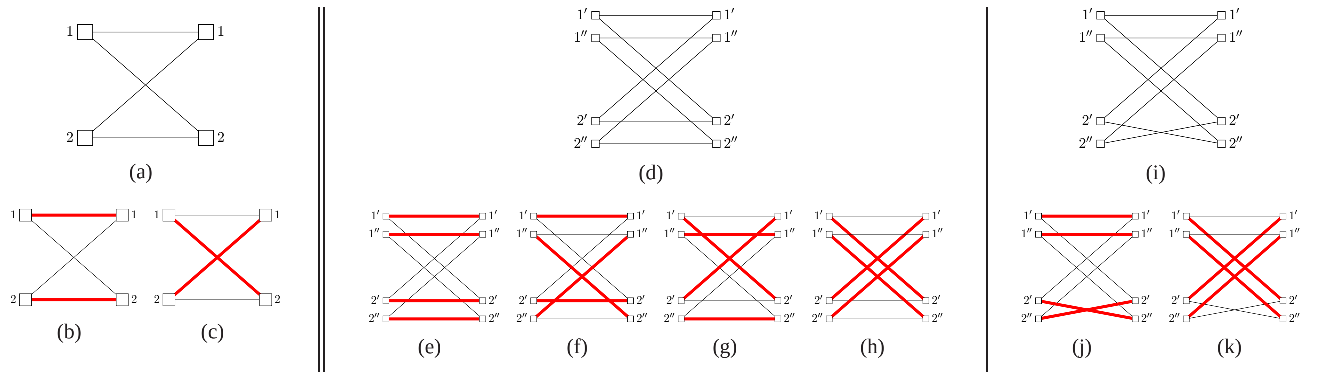


Fig. 5. Graphs (NFGs) that are discussed in Sections VI-B–VI-D. (a) Base graph. (b)–(c) Perfect matchings of the graph in (a). (d) A possible double cover of the graph in (a). (e)–(h) Perfect matchings of the graph in (d). (i) A possible double cover of the graph in (a). (j)–(k) Perfect matchings of the graph in (i).

- For $\theta_{1,1}\theta_{2,2} < \theta_{1,2}\theta_{2,1}$ the minimum is achieved at $\boldsymbol{\gamma} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$.

$\square$

Example 41 *For $n = 2$ and $\theta_{i,j} = 1$, $(i,j) \in \mathcal{I} \times \mathcal{J}$, we have*

$$\begin{aligned} \mathrm{perm}(\boldsymbol{\theta}) &= 2, \\ \mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) &= 1. \end{aligned}$$

Recall that $\mathrm{perm}(\boldsymbol{\theta})$ represents the sum of all the weighted perfect matchings of the complete bipartite graph $\mathsf{N}(\boldsymbol{\theta})$, and so, for the special choice $\theta_{i,j} = 1$, $(i,j) \in \mathcal{I} \times \mathcal{J}$, the quantity $\mathrm{perm}(\boldsymbol{\theta})$ represents the number of perfect matchings of $\mathsf{N}(\boldsymbol{\theta})$. As is illustrated in Figures 5(b)–(c), the graph $\mathsf{N}(\boldsymbol{\theta})$ has two perfect matchings, thereby combinatorially verifying $\mathrm{perm}(\boldsymbol{\theta}) = 2$. $\square$

B. The Degree-M Bethe Permanent for Matrices of Size 2×2 — Initial Considerations

One of the goals of this and the next subsections is to obtain a better combinatorial understanding of the result $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) = 1$ for $n = 2$, in particular, why it is different from $\mathrm{perm}(\boldsymbol{\theta})$, yet not too different.

Towards this goal, let us study the degree-M Bethe permanent of $\boldsymbol{\theta}$ as specified in Definition 38. Therein, the average is taken over $\left|\widetilde{\Psi}_M\right| = (M!)^4$ matrices

$$\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}} = \begin{pmatrix} \theta_{1,1}\widetilde{\boldsymbol{P}}_{1,1} & \theta_{1,2}\widetilde{\boldsymbol{P}}_{1,2} \\ \theta_{2,1}\widetilde{\boldsymbol{P}}_{2,1} & \theta_{2,2}\widetilde{\boldsymbol{P}}_{2,2} \end{pmatrix}, \quad \boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}} \in \widetilde{\Psi}_M.$$

We can simplify the analysis by realizing that the permanent of $\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}$ equals the permanent of a modified matrix $\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}$, where the first block row is multiplied from the left by $\widetilde{\boldsymbol{P}}_{1,1}^{-1}$, where the second block row is multiplied from the left by $\widetilde{\boldsymbol{P}}_{2,1}^{-1}$, and where the second block column is multiplied from the right by $\widetilde{\boldsymbol{P}}_{1,2}^{-1} \cdot \widetilde{\boldsymbol{P}}_{1,1}$, *i.e.*,

$$\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right) = \mathrm{perm}\begin{pmatrix} \theta_{1,1}\widetilde{\boldsymbol{I}} & \theta_{1,2}\widetilde{\boldsymbol{I}} \\ \theta_{2,1}\widetilde{\boldsymbol{I}} & \theta_{2,2}\widetilde{\boldsymbol{P}}_{2,1}^{-1}\widetilde{\boldsymbol{P}}_{2,2}\widetilde{\boldsymbol{P}}_{1,2}^{-1}\widetilde{\boldsymbol{P}}_{1,1} \end{pmatrix},$$

where $\widetilde{\boldsymbol{I}}$ is the identity matrix of size $M \times M$. Therefore, we can rewrite $\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta})$ as follows

$$\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \triangleq \sqrt[M]{\left\langle \mathrm{perm}\begin{pmatrix} \theta_{1,1}\widetilde{\boldsymbol{I}} & \theta_{2,1}\widetilde{\boldsymbol{I}} \\ \theta_{2,1}\widetilde{\boldsymbol{I}} & \theta_{2,2}\widetilde{\boldsymbol{P}}'_{2,2} \end{pmatrix} \right\rangle_{\widetilde{\boldsymbol{P}}'_{2,2} \in \mathcal{P}_{M \times M}}}, \tag{9}$$

i.e., an average over the $M!$ permutation matrices of size $M \times M$.

C. The Degree-M Bethe Permanent for Matrices of Size 2×2 — All-One Matrix

In this subsection we consider the cases $M = 2$, $M = 3$, and general M for the special choice

$$\boldsymbol{\theta} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Example 42 *Let $n = 2$, $M = 2$, and $\theta_{i,j} = 1$, $(i,j) \in \mathcal{I} \times \mathcal{J}$. We make the following observations.*

- *The average in* (9) *is over $2! = 2$ matrices, namely over*

$$\boldsymbol{\theta}^{\uparrow(1)} \triangleq \left(\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right), \quad \boldsymbol{\theta}^{\uparrow(2)} \triangleq \left(\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right).$$

- *The matrix $\boldsymbol{\theta}^{\uparrow(1)}$ corresponds to the double cover of $\mathsf{N}(\boldsymbol{\theta})$ shown in Figure 5(d). Because that graph has 4 perfect matchings,* cf. *Figures 5(e)–(h), we have*

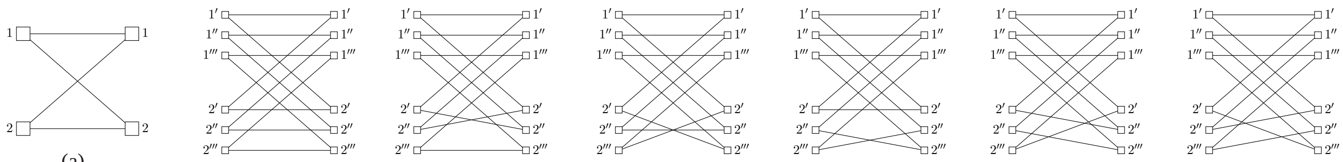
$$\mathrm{perm}(\boldsymbol{\theta}^{\uparrow(1)}) = 4.$$

- *The matrix $\boldsymbol{\theta}^{\uparrow(2)}$ corresponds to the double cover of $\mathsf{N}(\boldsymbol{\theta})$ shown in Figure 5(i). Because that graph has 2 perfect matchings,* cf. *Figures 5(j)–(k), we have*

$$\mathrm{perm}(\boldsymbol{\theta}^{\uparrow(1)}) = 2.$$

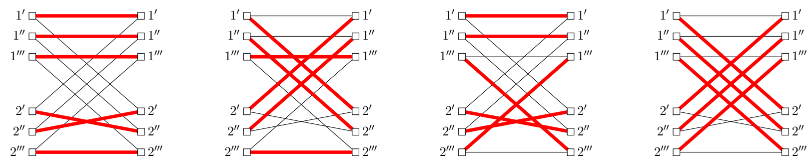
Putting everything together, we obtain the degree-2 Bethe permanent of $\boldsymbol{\theta}$, i.e.,

$$\mathrm{perm}_{\mathrm{B},2}(\boldsymbol{\theta}) = \sqrt[2]{\frac{1}{2!}\cdot(4+2)} = \sqrt[2]{\frac{1}{2!}\cdot 6} = \sqrt[2]{3} \approx 1.732.$$



(a) (b) 8 pms. (c) 4 pms. (d) 4 pms. (e) 4 pms. (f) 2 pms. (g) 2 pms.

Fig. 6. Graphs (NFGs) that are discussed in Sections VI-B–VI-D. (a) Base graph. (b)–(g) Possible triple covers of the graph in (a). ("pms." stands for "perfect matchings".)



(a) (b) (c) (d)

Fig. 7. The four perfect matchings of the triple cover in Figure 6(c).

*We note that the graph in Figure 5(d) consists of M independent copies of the graph in Figure 5(a), therefore it is not surprising that $\mathrm{perm}(\boldsymbol{\theta}^{\uparrow(1)}) = \mathrm{perm}(\boldsymbol{\theta})^M = 2^2 = 4$. On the other hand, the graph in Figure 5(d) consists of M **coupled** copies of the graph in Figure 5(a), which implies that we cannot choose the perfect matchings independently. Therefore, it is not surprising that we have $\mathrm{perm}(\boldsymbol{\theta}^{\uparrow(2)}) \neq \mathrm{perm}(\boldsymbol{\theta})^M = 2^2 = 4$, which finally results in $\mathrm{perm}_{\mathrm{B},2}(\boldsymbol{\theta}) \neq \mathrm{perm}(\boldsymbol{\theta})$. Nevertheless, these considerations also show why $\mathrm{perm}_{\mathrm{B},2}(\boldsymbol{\theta})$ is not too different from $\mathrm{perm}(\boldsymbol{\theta})$.* $\square$

Example 43 *Let $n = 2$, $M = 3$, and $\theta_{i,j} = 1$, $(i,j) \in \mathcal{I} \times \mathcal{J}$. The average in* (9) *is over $3! = 6$ matrices. These matrices correspond to the triple covers of $\mathsf{N}(\boldsymbol{\theta})$ shown in Figure 6(b)–(g). Computing the number of perfect matchings for each of these cases, we obtain*

$$\begin{aligned}\mathrm{perm}_{\mathrm{B},3}(\boldsymbol{\theta}) &= \sqrt[3]{\frac{1}{3!} \cdot (8+4+4+4+2+2)} \\ &= \sqrt[3]{\frac{1}{3!} \cdot 24} = \sqrt[3]{4} \approx 1.587.\end{aligned}$$

In particular, for the triple cover in Figure 6(c) we show its 4 perfect matchings explicitly in Figure 7.

*Overall, we can make similar observations as at the end of Example 42 concerning the **coupling** of the M copies of $\mathsf{N}(\boldsymbol{\theta})$ that make up a degree-M cover and its influence on the number of perfect matchings.* $\square$

Example 44 *Let $n = 2$, $M \in \mathbb{Z}_{>0}$, and $\theta_{i,j} = 1$, $(i,j) \in \mathcal{I} \times \mathcal{J}$. The average in* (9) *is over $M!$ matrices that correspond to the M-covers of $\mathsf{N}(\boldsymbol{\theta})$. For each of these matrices, their permanent equals the number of perfect matchings in the corresponding M-cover. We make the following observations (see Figures 5–7 for illustrations for the cases $M = 2$ and $M = 3$).*

- *Every M-cover consists of up to M cycles.*
- *Every cycle supports two perfect matchings (independently of the cycle length and independently of the perfect matchings chosen on the rest of the graph).*

Therefore, if an M-cover has c cycles then it has 2^c perfect matchings. The average in (9) *can then be evaluated with suitable combinatorial tools, for example by using the so-called cycle index of the symmetric group over M elements (see,* e.g., *[48]), and we obtain*

$$\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) = \sqrt[M]{M+1},$$

Therefore, in the limit $M \to \infty$, we obtain

$$\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) = \limsup_{M \to \infty} \ \mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) = 1.$$

This confirms the result for $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})$ in Example 41, which was obtained by analytical means. $\square$

D. The Degree-M Bethe Permanent for Matrices of Size 2×2 — General Non-Negative Matrix

In this subsection we consider the cases $M = 2$, $M = 3$, and general M for the general non-negative matrix

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{1,1} & \theta_{1,2} \\ \theta_{2,1} & \theta_{2,2} \end{pmatrix}.$$

A particular goal of this subsection is to compare the degree-M Bethe permanent of $\boldsymbol{\theta}$ with the permanent of $\boldsymbol{\theta}$. In fact, as we will see, for every considered case in this subsection we have $\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \leqslant \mathrm{perm}(\boldsymbol{\theta})$.

Example 45 *Let $n = 2$ and $M = 2$. We perform similar computations as in Example 42, but for a general non-negative matrix $\boldsymbol{\theta}$. Towards computing $\mathrm{perm}_{\mathrm{B},2}(\boldsymbol{\theta})$ as given in* (9)*, we make the following observations.*

- *The average in* (9) *is over $2! = 2$ matrices, namely over*

$$\boldsymbol{\theta}^{\uparrow(1)} \triangleq \left(\begin{array}{cc|cc} \theta_{1,1} & 0 & \theta_{1,2} & 0 \\ 0 & \theta_{1,1} & 0 & \theta_{1,2} \\ \hline \theta_{2,1} & 0 & \theta_{2,2} & 0 \\ 0 & \theta_{2,1} & 0 & \theta_{2,2} \end{array}\right),$$

$$\boldsymbol{\theta}^{\uparrow(2)} \triangleq \left(\begin{array}{cc|cc} \theta_{1,1} & 0 & \theta_{1,2} & 0 \\ 0 & \theta_{1,1} & 0 & \theta_{1,2} \\ \hline \theta_{2,1} & 0 & 0 & \theta_{2,2} \\ 0 & \theta_{2,1} & \theta_{2,2} & 0 \end{array}\right).$$

- *We obtain*

$$\begin{aligned}\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow(1)}\right) &= (\theta_{1,1}\theta_{2,2} + \theta_{1,2}\theta_{2,1})^2 \\ &= \theta_{1,1}^2\theta_{2,2}^2 + 2\theta_{1,1}\theta_{1,2}\theta_{2,1}\theta_{2,2} + \theta_{1,2}^2\theta_{2,1}^2.\end{aligned}$$

  Note that the coefficients add up to 4 because $\boldsymbol{\theta}^{\uparrow(1)}$ corresponds to the double cover of $\mathsf{N}(\boldsymbol{\theta})$ shown in Figure 5(d), which admits 4 (weighted) perfect matchings.
- *We obtain*

$$\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow(2)}\right) = \theta_{1,1}^2\theta_{2,2}^2 + \theta_{1,2}^2\theta_{2,1}^2.$$

  Note that the coefficients add up to 2 because $\boldsymbol{\theta}^{\uparrow(1)}$ corresponds to the double cover of $\mathsf{N}(\boldsymbol{\theta})$ shown in Figure 5(i), which admits 2 (weighted) perfect matchings.

Putting everything together, we obtain for the square of the degree-2 Bethe partition function of $\boldsymbol{\theta}$

$$\begin{aligned}\big(\mathrm{perm}_{\mathrm{B},2}(\boldsymbol{\theta})\big)^2 &= \frac{1}{2}\cdot\big(\mathrm{perm}(\boldsymbol{\theta}^{\uparrow(1)}) + \mathrm{perm}(\boldsymbol{\theta}^{\uparrow(2)})\big) \\ &= \theta_{1,1}^2\theta_{2,2}^2 + \theta_{1,1}\theta_{1,2}\theta_{2,1}\theta_{2,2} + \theta_{1,2}^2\theta_{2,1}^2.\end{aligned}$$

Given the observations that

$$\begin{aligned}\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow(1)}\right) &\leqslant \big(\mathrm{perm}(\boldsymbol{\theta})\big)^2, \\ \mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow(2)}\right) &\leqslant \big(\mathrm{perm}(\boldsymbol{\theta})\big)^2,\end{aligned}$$

it is not surprising that we also have the inequality

$$\big(\mathrm{perm}_{\mathrm{B},2}(\boldsymbol{\theta})\big)^2 \leqslant \big(\mathrm{perm}(\boldsymbol{\theta})\big)^2,$$

i.e.,

$$\mathrm{perm}_{\mathrm{B},2}(\boldsymbol{\theta}) \leqslant \mathrm{perm}(\boldsymbol{\theta}).$$

□

Example 46 *Let $n = 2$ and $M = 3$. We perform similar computations as in Example 43, but for a general non-negative matrix $\boldsymbol{\theta}$. Towards computing $\mathrm{perm}_{\mathrm{B},3}(\boldsymbol{\theta})$ as given in (9), we make the following observations.*

- *The average in (9) is over $3! = 6$ matrices. These matrices correspond to the triple covers of $\mathsf{N}(\boldsymbol{\theta})$ shown in Figure 6(b)–(g).*
- *For example, for the matrix $\boldsymbol{\theta}^{\uparrow(2)}$ corresponding to the triple cover in Figure 6(c), we obtain*

$$\begin{aligned}\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow(2)}\right) &= \theta_{1,1}^3\theta_{2,2}^3 + \theta_{1,1}^1\theta_{1,2}^2\theta_{2,1}^2\theta_{2,2}^1 \\ &\quad + \theta_{1,1}^2\theta_{1,2}^1\theta_{2,1}^1\theta_{2,2}^2 + \theta_{1,2}^3\theta_{2,1}^3,\end{aligned}$$

  where each (weighted) perfect matching in Figure 7 contributes one monomial to the above expression. One can verify that

$$\begin{aligned}\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow(2)}\right) &= \big(\theta_{1,1}^2\theta_{2,2}^2+\theta_{1,2}^2\theta_{2,1}^2\big)\cdot(\theta_{1,1}\theta_{2,2}+\theta_{1,2}\theta_{2,1}) \\ &\leqslant \big(\theta_{1,1}\theta_{2,2}+\theta_{1,2}\theta_{2,1}\big)^2\cdot(\theta_{1,1}\theta_{2,2}+\theta_{1,2}\theta_{2,1}) \\ &= \big(\theta_{1,1}\theta_{2,2}+\theta_{1,2}\theta_{2,1}\big)^3 \\ &= \big(\mathrm{perm}(\boldsymbol{\theta})\big)^3.\end{aligned}$$

  (The product expression in the first line is not surprising given the fact that graph in Figure 6(c) contains two independent components, each contributing one factor to the above product.)

Similar observations can be made for the other five triple covers in Figure 6(b)–(g), and so we obtain

$$\big(\mathrm{perm}_{\mathrm{B},3}(\boldsymbol{\theta})\big)^3 \leqslant \big(\mathrm{perm}(\boldsymbol{\theta})\big)^3,$$

i.e.,

$$\mathrm{perm}_{\mathrm{B},3}(\boldsymbol{\theta}) \leqslant \mathrm{perm}(\boldsymbol{\theta}).$$

□

Example 47 *Let $n = 2$ and $M \in \mathbb{Z}_{>0}$. We perform similar computations as in Example 44, but for a general non-negative matrix $\boldsymbol{\theta}$. The observations that we made there can be generalized (beyond the all-one matrix), and we obtain*

$$\big(\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta})\big)^M = \sum_{\ell=0}^{M}(\theta_{1,1}\theta_{2,2})^{M-\ell}(\theta_{1,2}\theta_{2,1})^{\ell}.$$

Because

$$\big(\mathrm{perm}(\boldsymbol{\theta})\big)^M = \sum_{\ell=0}^{M}\binom{M}{\ell}(\theta_{1,1}\theta_{2,2})^{M-\ell}(\theta_{1,2}\theta_{2,1})^{\ell},$$

we see that

$$\big(\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta})\big)^M \leqslant \big(\mathrm{perm}(\boldsymbol{\theta})\big)^M,$$

i.e.,

$$\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \leqslant \mathrm{perm}(\boldsymbol{\theta}).$$

Moreover, in the limit $M \to \infty$, we have

$$\begin{aligned}\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) &= \limsup_{M\to\infty}\ \mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \\ &= \max(\theta_{1,1}\theta_{2,2},\ \theta_{2,1}\theta_{1,2}).\end{aligned}$$

This confirms the result for $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})$ in Lemma 40, which was obtained by analytical means. □

For $n > 2$, we leave it as an open problem to obtain an explicit expressions for $\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta})$, $M \in \mathbb{Z}_{>0}$, either for the all-one matrix case, or for the general non-negative matrix case.

In conclusion, the above examples shows that in general $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) \neq \mathrm{perm}(\boldsymbol{\theta})$, however, they also show that the Bethe permanent has the potential to give reasonably good estimates, in particular in the cases where the "coupling effect" in the average graph cover is not too strong. Heuristically, this "coupling effect" seems actually to be the worst for $n = 2$ and become weaker the larger n is.

E. Relevance of Finite Graph Covers

If the NFG $\mathsf{N}(\boldsymbol{\theta})$ had no cycles then the SPA could be used to exactly compute the partition function. Namely, after a finite number of iterations, the SPA would reach a fixed point and the partition function $Z_{\mathrm{G}}\big(\mathsf{N}(\boldsymbol{\theta})\big) = \mathrm{perm}(\boldsymbol{\theta})$ could be computed with the help of an expression like $\exp\big(-F_{\mathrm{Bethe}}^{\#}\big(\{\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\},\{\overrightarrow{\mathrm{V}}_{i,j}^{(t)}\}\big)\big)$, where $F_{\mathrm{Bethe}}^{\#}$ is defined in Lemma 31. However, $\mathsf{N}(\boldsymbol{\theta})$ has cycles: the use of this

expression at a fixed point of the SPA is still possible but usually it does not yield the correct partition function. In this subsection, we would like to better understand the source of this suboptimality.

To that end, observe that the SPA is an algorithm that processes information locally on $\mathsf{N}(\boldsymbol{\theta})$, *i.e.*, messages are sent along edges, function nodes take incoming messages from incident edges, do some computations, and send out new messages along the incident edges. On the one hand, this locality explains the main strengths of the SPA, namely its low complexity and its parallelizability, two key factors for making the SPA a popular algorithm. On the other hand, this locality explains also the main weakness of the SPA. Namely, a locally operating like SPA "cannot distinguish" if it is operating on $\mathsf{N}(\boldsymbol{\theta})$ or any of its covers [11], [49], [50].

More precisely, let $\widetilde{\mathsf{N}}$ be an M-cover $\widetilde{\mathsf{N}}$ of $\mathsf{N}(\boldsymbol{\theta})$. Such an M-cover "looks locally the same" as $\mathsf{N}(\boldsymbol{\theta})$ in the sense that the local structure of $\widetilde{\mathsf{N}}$ is exactly the same as the one of $\mathsf{N}(\boldsymbol{\theta})$. (Of course, globally $\widetilde{\mathsf{N}}$ and $\mathsf{N}(\boldsymbol{\theta})$ are different because the former NFG contains M times as many function nodes and M times as many edges.) Consequently, if the SPA is run on $\widetilde{\mathsf{N}}$ with the same initialization as the SPA on $\mathsf{N}(\boldsymbol{\theta})$ (every initial message is replicated M times), we observe that, because both graphs look locally the same and because the SPA is a locally operating algorithm, after every iteration the messages on $\widetilde{\mathsf{N}}$ are exactly the same as the messages on $\mathsf{N}(\boldsymbol{\theta})$, simply replicated M times. In that sense, the SPA "cannot distinguish" if it is operating on $\mathsf{N}(\boldsymbol{\theta})$ or $\widetilde{\mathsf{N}}$, or, in fact, any other M-cover of $\mathsf{N}(\boldsymbol{\theta})$. This observation allows us to give the following interpretation of (8) (which is reproduced here for the ease of reference)

$$\mathrm{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \triangleq \sqrt[M]{\left\langle \mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow\tilde{\boldsymbol{P}}}\right)\right\rangle_{\tilde{\boldsymbol{P}}\in\tilde{\Psi}_M}}. \tag{10}$$

Namely, because the SPA implicitly tries to compute in parallel the partition function $Z_{\mathrm{G}}\big(\mathsf{N}(\boldsymbol{\theta})^{\uparrow\tilde{\boldsymbol{P}}}\big) = \mathrm{perm}(\boldsymbol{\theta}^{\uparrow\tilde{\boldsymbol{P}}})$ for all M-covers of $\mathsf{N}(\boldsymbol{\theta})$, yet it has to give back one real number only, the "best it can do" is to give back the average of these partition functions, *i.e.*, $\left\langle \mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow\tilde{\boldsymbol{P}}}\right)\right\rangle_{\tilde{\boldsymbol{P}}\in\tilde{\Psi}_M}$. (The M-th root that appears in (10) is included so that the result is properly normalized w.r.t. $Z_{\mathrm{G}}\big(\mathsf{N}(\boldsymbol{\theta})\big) = \mathrm{perm}(\boldsymbol{\theta})$.)

Let us conclude this section with a comment on a paper by Barvinok [23] that presents bounds on the number of zero/one matrices with prescribed row and column sums. (As already mentioned in Section I-C, in statistical physics terms the approach taken therein can be considered as a mean-field approach.) In terms of NFGs, the quantity of interest is expressed as the partition function of an NFG that has the same topology as $\mathsf{N}(\boldsymbol{\theta})$ but different function nodes.

Section 3.1 of [23] then presents an interpretation of these bounds that has a similar flavor of the graph cover interpretation of the Bethe permanent, however, it also has stark differences. Namely, in terms of NFGs, Section 3.1 of [23] presents an NFG where every function node of the base graph is replicated M times and every edge is replicated M^2 times, *i.e.*, all Mn left-hand side function nodes are connected by exactly one edge to all the Mn right-hand side function nodes. In order for this to make sense, the local functions are adapted so that they have Mn arguments instead of n arguments. It is then shown that the M^2-th root of the partition function of this new NFG, $M \to \infty$, yields the relevant number in which the bounds are expressed. Despite all the similarities, the differences to finite graph covers are clear:

- There is only one such M-fold version of the base graph, whereas the number of M-covers of $\mathsf{N}(\boldsymbol{\theta})$ is $(M!)^{(n^2)}$.
- The number of edges is M^2n^2, whereas the number of edges in an M-cover of $\mathsf{N}(\boldsymbol{\theta})$ is Mn^2.
- The local functions need to be adapted in order to allow for Mn instead of n arguments, whereas the local functions of an M-cover of $\mathsf{N}(\boldsymbol{\theta})$ are the same as the local functions of $\mathsf{N}(\boldsymbol{\theta})$.

VII. The Relationship between the Permanent and the Bethe Permanent

In this section we explore the relationship between $\mathrm{perm}(\boldsymbol{\theta})$ and $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})$, in particular, if and how the $\mathrm{perm}(\boldsymbol{\theta})$ can be upper and lower bounded by expressions that are functions of $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})$. For an additional/complementary discussion on this topic we refer to [17].

We start with a lemma that shows that there are non-negative square matrices for which the Bethe permanent can give rather accurate estimates of the permanent, thereby showing the overall potential of the Bethe permanent to be the basis for good upper and lower bounds on the permanent of general non-negative square matrix.

Lemma 48 *Let $\mathbf{1}_{n\times n}$ be the all-one matrix of size $n\times n$. Then*

$$\frac{\mathrm{perm}(\mathbf{1}_{n\times n})}{\mathrm{perm}_{\mathrm{B}}(\mathbf{1}_{n\times n})} = \sqrt{\frac{2\pi n}{\mathrm{e}}}\cdot\big(1+o(1)\big),$$

where $o(1)$ is w.r.t. n.

Proof: See Appendix H. □

Although the factor $\sqrt{2\pi n/\mathrm{e}}$ is non-negligible, compared to $\mathrm{perm}(\mathbf{1}_{n\times n}) = n!$ it is rather small.

A. Lower Bounds on the Permanent of the Matrix $\boldsymbol{\theta}$

In this subsection we study lower bounds on $\mathrm{perm}(\boldsymbol{\theta})$ based on $\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})$.

Theorem 49 (Gurvits [10]) *It holds that*

$$\frac{\mathrm{perm}(\boldsymbol{\theta})}{\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})} \geqslant 1.$$

Proof: This result was recently shown by Gurvits [10]. Roughly speaking, its elegant proof is based on first expressing $\boldsymbol{\theta}$ in terms of a stationary point of $F_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}$ and then applying an inequality due to Schrijver [51]. For more details, along with a discussion of this result's relationship to the results in [52], [53], we refer to [10]. □

Corollary 50 (Gurvits [10]) *For any $\boldsymbol{\gamma}\in\Gamma_{n\times n}$ it holds that*

$$\frac{\mathrm{perm}(\boldsymbol{\theta})}{\exp\big(-F_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}(\boldsymbol{\gamma})\big)} \geqslant 1.$$

Proof: This is a straightforward consequence of Theorem 49 and Definitions 11 and 12. $\square$

This corollary has its significance when one is not willing to run the SPA algorithm, but one has a reasonably good estimate of the $\boldsymbol{\gamma} \in \Gamma_{n\times n}$ that minimizes $F_{\mathrm{B,N}(\boldsymbol{\theta})}$. This approach is for example interesting when one wants to obtain analytical lower bounds on the permanent of some parametrized class of non-negative square matrices.

In the Allerton 2010 version of this paper we also stated the inequality that appears in Theorem 49. However, while writing the present paper we realized that our "proof" of that theorem had a flaw, which, so far, we have not been able to fix. However, we still think that our proof strategy can work out and possibly give an alternative viewpoint of Schrijver's inequality that features prominently in [10]. In that respect, we list below some special cases of matrices $\boldsymbol{\theta}$ for which our proof strategy works, along with conjectures that, if true, would give an alternative proof of Theorem 49 in its full generality.

Conjecture 51 *For any $M \in \mathbb{Z}_{>0}$ it holds that*

$$\left\langle \operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right)\right\rangle_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}_M} \leqslant \big(\operatorname{perm}(\boldsymbol{\theta})\big)^M.$$

Possibly also the following, stronger, statement is true: for any $M \in \mathbb{Z}_{>0}$ and any $\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M$ it holds that

$$\operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right) \leqslant \big(\operatorname{perm}(\boldsymbol{\theta})\big)^M.$$

$\square$

Theorem 49 would then follow from

$$\begin{aligned}
\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta}) &\overset{\text{(a)}}{=} \limsup_{M\to\infty} \ \operatorname{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) \\
&\overset{\text{(b)}}{=} \limsup_{M\to\infty} \ \sqrt[M]{\left\langle \operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right)\right\rangle_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}_M}} \\
&\overset{\text{(c)}}{\leqslant} \limsup_{M\to\infty} \ \sqrt[M]{\operatorname{perm}(\boldsymbol{\theta})^M} \\
&= \limsup_{M\to\infty} \ \operatorname{perm}(\boldsymbol{\theta}) \\
&\overset{\text{(d)}}{=} \operatorname{perm}(\boldsymbol{\theta}),
\end{aligned}$$

where at step (a) we have used Theorem 39, where at step (b) we have used Definition 38, where at step (c) we have used the weaker part of Conjecture 51, and where step (d) follows from evaluating the (now trivial) limit $M \to \infty$.

We now list some special matrices $\boldsymbol{\theta}$ for which Conjecture 51 is true.

- Conjecture 51 is true for $\boldsymbol{\theta} = \mathbf{1}_{n\times n}$. (The proof is given in Appendix I.)
- Conjecture 51 is true for all matrices $\boldsymbol{\theta}$ that were studied in Section VI.

Actually, the results in Section VI suggest the following, stronger version of Conjecture 51.

Conjecture 52 *Fix some $M \in \mathbb{Z}_{>0}$ and consider the expressions*

$$\left\langle \operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right)\right\rangle_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}_M} \quad \textit{and} \quad \big(\operatorname{perm}(\boldsymbol{\theta})\big)^M$$

as polynomials in the indeterminates $\{\theta_{i,j}\}_{i,j}$. We conjecture that the coefficient of every monomial of the first polynomial is upper bounded by the coefficient of the corresponding monomial of the second polynomial.

Possibly also the following, stronger, statement is true. Fix some $M \in \mathbb{Z}_{>0}$ and $\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M$, and consider the expressions

$$\operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right) \quad \textit{and} \quad \big(\operatorname{perm}(\boldsymbol{\theta})\big)^M$$

as polynomials in the indeterminates $\{\theta_{i,j}\}_{i,j}$. We conjecture that the coefficient of every monomial of the first polynomial is upper bounded by the coefficient of the corresponding monomial of the second polynomial. $\square$

B. Upper Bounds on the Permanent of the Matrix $\boldsymbol{\theta}$

In this subsection we list conjectures and open problems w.r.t. upper bounds on $\operatorname{perm}(\boldsymbol{\theta})$ based on $\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})$.

Conjecture 53 (Gurvits [10]) *Let $\boldsymbol{\theta}$ be an arbitrary non-negative matrix of size $n \times n$. For even n it is conjectured that*

$$\frac{\operatorname{perm}(\boldsymbol{\theta})}{\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})} \leqslant \sqrt{2}^{\,n}, \tag{11}$$

with a similar conjecture for odd n. Note that (11) *holds with equality for the matrix $\boldsymbol{\theta} = \boldsymbol{I}_{(n/2)\times(n/2)} \otimes \mathbf{1}_{2\times 2}$, which is the Kronecker product of an identity matrix of size $(n/2) \times (n/2)$ and the all-one matrix of size 2×2.* $\square$

The above conjecture replaces the conjecture that we made in the Allerton 2010 version of this paper where, for fixed n, the largest ratio $\operatorname{perm}(\boldsymbol{\theta})/\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})$ was thought to be obtained for the all-one matrix of size $n \times n$.

Besides proving the bound in Conjecture 53, it would be desirable to prove statements of the form

$$\Pr\left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \frac{\operatorname{perm}(\boldsymbol{\theta})}{\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})} \leqslant \tau\right\} \geqslant 1 - \varepsilon,$$

where $\boldsymbol{\Theta}$ is some ensemble of random matrices of size $n \times n$, where τ is some positive real number, and where ε is some small positive number. For example, for the ensemble of $n \times n$ matrices where the matrix entries are chosen uniformly and independently between 0 and 1, we conjecture that $\operatorname{perm}(\boldsymbol{\theta})/\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})$ is, with high probability, upper bounded by the ratio that appears in Lemma 48. (Note that this ratio is much smaller than the ratio that appears in Conjecture 53.)

C. Closeness of the Permanent to the Bethe Permanent

In this subsection we list some cases where $\operatorname{perm}(\boldsymbol{\theta})$ is relatively close to $\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})$. We start with an auxiliary result that relates the Bethe permanent of a lifted matrix to the Bethe permanent of the base matrix.

Lemma 54 *For any $M \in \mathbb{Z}_{>0}$ and any $\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M$ it holds that*

$$\operatorname{perm}_{\mathrm{B}}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right) = \big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M$$

Proof: See Appendix J. $\square$

Theorem 55 *For any $\alpha > 1$ and any $M \geqslant M_\alpha$, the majority of the matrices $\left\{\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right\}_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}_M}$ satisfies*

$$1 \leqslant \frac{\operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)}{\operatorname{perm}_{\mathrm{B}}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)} < \alpha^M$$

Here M_α is a parameter that depends on α.

Proof: The first inequality follows from Theorem 49. We prove the second inequality by contradiction. So, assume that there is an $\alpha > 1$ and a constant M_α such that for all $M \geqslant M_\alpha$ the set $\widetilde{\Psi}'_M \subseteq \widetilde{\Psi}_M$ of all lifted matrices $\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}$ that satisfy $\operatorname{perm}\big(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\big) \geqslant \alpha^M \cdot \operatorname{perm}_{\mathrm{B}}\big(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\big)$ has size at least $|\widetilde{\Psi}_M|/2$. Then

$$\begin{aligned}
\operatorname{perm}_{\mathrm{B},M}(\boldsymbol{\theta}) &\overset{\text{(a)}}{=} \sqrt[M]{\left\langle \operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)\right\rangle_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}_M}} \\
&\overset{\text{(b)}}{=} \sqrt[M]{\frac{1}{|\widetilde{\Psi}_M|}\sum_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}_M} \operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)} \\
&\geqslant \sqrt[M]{\frac{1}{|\widetilde{\Psi}_M|}\sum_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}'_M} \operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)} \\
&\overset{\text{(c)}}{\geqslant} \sqrt[M]{\frac{1}{|\widetilde{\Psi}_M|}\sum_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}'_M} \alpha^M \cdot \operatorname{perm}_{\mathrm{B}}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)} \\
&\overset{\text{(d)}}{=} \sqrt[M]{\frac{1}{|\widetilde{\Psi}_M|}\sum_{\widetilde{\boldsymbol{P}}\in\widetilde{\Psi}'_M} \alpha^M \cdot \big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M} \\
&= \sqrt[M]{\frac{|\widetilde{\Psi}'_M|}{|\widetilde{\Psi}_M|}} \cdot \alpha \cdot \operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta}) \\
&\overset{\text{(e)}}{\geqslant} 2^{-1/M} \cdot \alpha \cdot \operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta}),
\end{aligned}$$

where at step (a) we have used Definition 38, where at step (b) we have replaced the angular brackets by the corresponding normalized sum, where at step (c) we have used the assumption, where at step (d) we have used Lemma 54, and where at step (e) we have again used the assumption. However, taking $\limsup_{M\to\infty}$ on both sides of the above expression, we see that we obtain a contradiction w.r.t. Theorem 39. $\square$

The following example partially corroborates Theorem 55.

Example 56 *For some positive integer M, consider the matrix*

$$\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}} = \begin{pmatrix} \theta_{1,1}\widetilde{\boldsymbol{I}} & \theta_{1,2}\widetilde{\boldsymbol{I}} \\ \theta_{2,1}\widetilde{\boldsymbol{I}} & \theta_{2,2}\widetilde{\boldsymbol{P}}'_{2,2} \end{pmatrix},$$

where $\widetilde{\boldsymbol{I}}$ is the identity matrix of size $M \times M$ and where $\widetilde{\boldsymbol{P}}'_{2,2}$ is a once cyclically left-shifted identity matrix of size $M \times M$. Then

$$\begin{aligned}
\operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right) &= \theta_{1,1}^M\theta_{2,2}^M + \theta_{1,2}^M\theta_{2,1}^M, \\
\big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M &= \big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M \\
&= \big(\max(\theta_{1,1}\theta_{2,2},\, \theta_{1,2}\theta_{2,1})\big)^M,
\end{aligned}$$

where the first result is a consequence of the observation that the underlying graph has exactly one cycle, i.e., *only two perfect matchings, and where the second result follows from Lemmas 40 and 54. Therefore,*

$$1 \leqslant \frac{\operatorname{perm}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)}{\operatorname{perm}_{\mathrm{B}}\left(\boldsymbol{\theta}^{\uparrow\widetilde{\boldsymbol{P}}}\right)} \leqslant 2.$$

Note that the right-hand side of the above expression does not only grow sub-exponentially in M, it does not grow at all. $\square$

Let us conclude this subsection with the following remark. As already mentioned, the proof of Theorem 49 takes advantage of an inequality by Schrijver [51], and therefore the closeness of $\operatorname{perm}(\boldsymbol{\theta})$ to $\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})$ is linked with the tightness of Schrijver's inequality. Now, interestingly enough, when Schrijver demonstrates a certain asymptotic tightness of his inequality, *cf.* [51, Section 3], he *implicitly* evaluates and compares both sides of his inequality for some finite cover of a certain graph.

D. Open Problems on the Relationship between the Permanent and the Bethe Permanent

There are also classes of structured matrices for which it would be interesting to better understand the relationship between the permanent and the Bethe permanent. For example, the permanent of the matrix

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha_1^{\mu_1} & \alpha_1^{\mu_2} & \cdots & \alpha_1^{\mu_m} & 1 & \cdots & 1 \\ \alpha_2^{\mu_1} & \alpha_2^{\mu_2} & \cdots & \alpha_2^{\mu_m} & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \alpha_n^{\mu_1} & \alpha_n^{\mu_2} & \cdots & \alpha_n^{\mu_m} & 1 & \cdots & 1 \end{pmatrix},$$

with $0 \leqslant m \leqslant n$, real numbers $\alpha_\ell \geqslant 0$, $\ell \in [n]$, and real numbers μ_ℓ, $\ell \in [m]$, turns up in a variety of contexts.

- When $\sum_{\ell\in[n]} \alpha_\ell = 1$ and μ_ℓ are non-negative integers then $\operatorname{perm}(\boldsymbol{\theta})$ corresponds to the probability of the pattern of a sequence (see, *e.g.*, [54]).
- When $m = n$ and $\mu_\ell = n - 1 - \ell$, $\ell \in [n]$, then $\operatorname{perm}(\boldsymbol{\theta})$ appears in the analysis of list ordering algorithms (see, *e.g.*, [55]) or in the analysis of source coding algorithms (see, *e.g.*, [56]). Note that in this case, $\boldsymbol{\theta}$ is a Vandermonde matrix.

Moreover, given the fact that the above $\boldsymbol{\theta}$ depends only on (at most) $2n$ parameters (and not on n^2 parameters as $\boldsymbol{\theta}$ in (1)), one wonders if speed-ups in the SPA-based computation of $\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})$ are possible.

Moreover, in some applications one is not interested in the absolute value of the permanent, only the relative value in the sense that for two matrices $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ one wants to know which one has the larger permanent. Therefore, for some suitable stochastic setting it would be desirable to state with what probability $\operatorname{perm}(\boldsymbol{\theta}) \leqslant \operatorname{perm}(\boldsymbol{\theta}')$ is equivalent to $\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta}) \leqslant \operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta}')$. Some very encouraging initial investigations of this topic have been presented in [8, Section 4.2].

VIII. Fractional Bethe Permanent

The terms that appear in $H_{\mathrm{B}}(\boldsymbol{\gamma})$ in Lemma 14 all have either coefficient $+1$ or -1. The main idea behind the fractional Bethe entropy function is to allow these coefficients to take on also other values. This is done towards the goal of obtaining a modified Bethe free energy function whose minimum resembles the minimum of the Gibbs free energy function even more.[13] Such generalizations of the Bethe entropy function were for example considered in [57]–[62] and a combinatorial characterization of the fractional Bethe entropy function was discussed in [45]. In particular, for the permanent estimation problem such generalizations are extensively studied in the very recent paper by A. B. Yedidia and Chertkov [17], to which we refer for additional discussion on this topic.

As we will see in this section, if the modifications to the Bethe entropy function are applied within some suitable limits, the concavity of the modified Bethe entropy function (and therefore the convexity of the modified Bethe free energy function) will be maintained.

Definition 57 *Let*

$$\boldsymbol{\kappa} \triangleq \Big\{ \{\kappa_i\}_{i\in\mathcal{I}}, \{\kappa_j\}_{j\in\mathcal{J}}, \{\kappa_{i,j}\}_{(i,)\in\mathcal{I}\times\mathcal{J}} \Big\}$$

be a collection of real values. We define the $\boldsymbol{\kappa}$-fractional Bethe entropy function to be

$$\begin{aligned} H_{\mathrm{B}}^{(\boldsymbol{\kappa})} &: \Gamma_{n\times n} \to \mathbb{R}, \\ \boldsymbol{\gamma} &\mapsto \sum_i \kappa_i \cdot H_{\mathrm{B},i}(\boldsymbol{\gamma}_i) + \sum_j \kappa_j \cdot H_{\mathrm{B},j}(\boldsymbol{\gamma}_i) \\ &\qquad - \sum_{i,j} \kappa_{i,j} \cdot H_{\mathrm{B},(i,j)}(\gamma_{i,j}). \end{aligned}$$

(Clearly, if all values in $\boldsymbol{\kappa}$ equal 1 then $H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma}) = H_{\mathrm{B}}(\boldsymbol{\gamma})$, with $H_{\mathrm{B}}(\boldsymbol{\gamma})$ as shown in Lemma 14.) □

Lemma 58 *The fractional Bethe entropy function from Definition 57 can also be expressed as follows*

$$\begin{aligned} H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma}) = &-\sum_{i,j} (\kappa_i + \kappa_j - \kappa_{i,j}) \cdot \gamma_{i,j} \log(\gamma_{i,j}) \\ &+ \sum_{i,j} \kappa_{i,j} \cdot (1-\gamma_{i,j}) \log(1-\gamma_{i,j}). \end{aligned}$$

(If all values in $\boldsymbol{\kappa}$ equal 1 then $H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma}) = H_{\mathrm{B}}(\boldsymbol{\gamma})$, with $H_{\mathrm{B}}(\boldsymbol{\gamma})$ as shown in Corollary 15.)

Proof: Follows from combining Definition 57 and Lemma 14. □

The following definition generalizes Definitions 11 and 12 and Corollary 15.

Definition 59 *We define the $\boldsymbol{\kappa}$-fractional Bethe free energy function to be*

$$\begin{aligned} F_{\mathrm{B}}^{(\boldsymbol{\kappa})} &: \Gamma_{n\times n} \to \mathbb{R}, \\ \boldsymbol{\gamma} &\mapsto U_{\mathrm{B}}(\boldsymbol{\gamma}) - H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma}), \end{aligned}$$

and the $\boldsymbol{\kappa}$-fractional Bethe permanent to be

$$\mathrm{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\theta}) \triangleq \exp\left(-\min_{\boldsymbol{\beta}\in\mathcal{B}} F_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\beta}) \right).$$

□

The following theorem gives a sufficient condition on $\boldsymbol{\kappa}$ so that the $\boldsymbol{\kappa}$-fractional Bethe entropy function is concave in $\boldsymbol{\gamma}$, thereby generalizing Theorem 22.

Theorem 60 *If $\boldsymbol{\kappa}$ is such that*

$$\begin{aligned} \kappa_i &\geqslant 0 && (i \in \mathcal{I}), \\ \kappa_j &\geqslant 0 && (j \in \mathcal{J}), \\ \kappa_i + \kappa_j &\geqslant 2\kappa_{i,j} && ((i,j) \in \mathcal{I}\times\mathcal{J}). \end{aligned}$$

then $H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ is a concave function of $\boldsymbol{\gamma}$ and $F_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ is a convex function of $\boldsymbol{\gamma}$.

Proof: We have

$$\begin{aligned} &H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma}) \\ &\overset{\text{(a)}}{=} -\sum_{i,j} \left(\frac{\kappa_i + \kappa_j}{2} + \frac{\kappa_i + \kappa_j}{2} - \kappa_{i,j} \right) \cdot \gamma_{i,j} \log(\gamma_{i,j}) \\ &\qquad + \sum_{i,j} \left(\frac{\kappa_i + \kappa_j}{2} - \frac{\kappa_i + \kappa_j}{2} + \kappa_{i,j} \right) \cdot (1-\gamma_{i,j}) \log(1-\gamma_{i,j}) \\ &\overset{\text{(b)}}{=} \sum_i \frac{\kappa_i}{2} \cdot S(\boldsymbol{\gamma}_i) + \sum_j \frac{\kappa_j}{2} \cdot S(\boldsymbol{\gamma}_j) \\ &\qquad + \sum_{i,j} \left(\frac{\kappa_i + \kappa_j}{2} - \kappa_{i,j} \right) \cdot h(\gamma_{i,j}), \end{aligned}$$

where at step (a) we have used Lemma 58, and where at step (b) we have used the S-function as specified in Definition 19 and have introduced the binary entropy function $h: [0,1] \to \mathbb{R},\ \xi \mapsto -\xi\log(\xi) - (1-\xi)\log(1-\xi)$. If $\kappa_i \geqslant 0$, $\kappa_j \geqslant 0$, and $\frac{\kappa_i+\kappa_j}{2} - \kappa_{i,j} \geqslant 0$ (the latter being equivalent to $\kappa_i + \kappa_j \geqslant 2\kappa_{i,j}$), then the concavity of $H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ in $\boldsymbol{\gamma}$ follows from Theorem 20, the well-known concavity of the binary entropy function, and the fact that the sum of concave functions is a concave function.

The convexity of $F_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ in $\boldsymbol{\gamma}$ follows from the concavity of $H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ in $\boldsymbol{\gamma}$ and the linearity of $U_{\mathrm{B}}(\boldsymbol{\gamma})$ in $\boldsymbol{\gamma}$. □

Lemma 61 *An interesting choice for $\boldsymbol{\kappa}$ is*

$$\begin{aligned} \kappa_i &= 1 && (i \in \mathcal{I}), \\ \kappa_j &= 1 && (j \in \mathcal{J}), \\ \kappa_{i,j} &= 1 - \frac{1}{2n} && ((i,j) \in \mathcal{I}\times\mathcal{J}). \end{aligned}$$

The resulting $H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ is a concave function of $\boldsymbol{\gamma}$ and the resulting $F_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ is a convex function of $\boldsymbol{\gamma}$. Moreover, letting $\mathbf{1}_{n\times n}$ be the all-one matrix of size $n\times n$, we obtain

$$\frac{\mathrm{perm}(\mathbf{1}_{n\times n})}{\mathrm{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\mathbf{1}_{n\times n})} = \frac{\sqrt{2\pi}}{\mathrm{e}} \cdot \big(1 + o(1)\big) = 0.922\ldots \cdot \big(1 + o(1)\big).$$

[13]One might also modify $U_{\mathrm{B}}(\boldsymbol{\gamma})$, however, we do not pursue this option here.

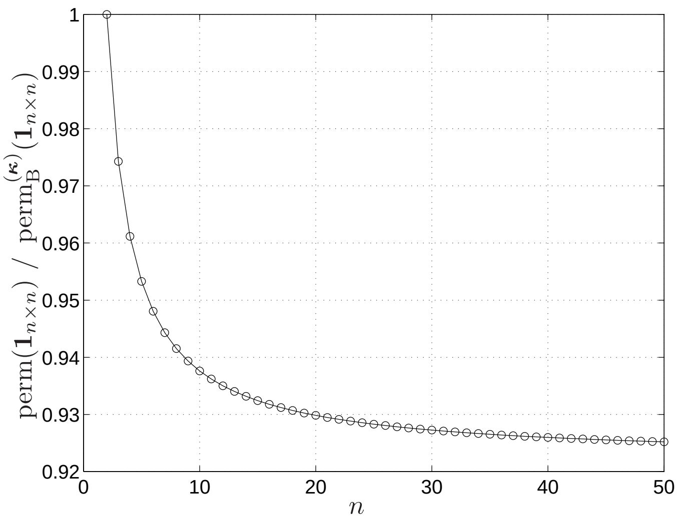


Fig. 8. Illustration of the ratio $\mathrm{perm}(\mathbf{1}_{n\times n})/\,\mathrm{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\mathbf{1}_{n\times n})$ for the special choice of $\boldsymbol{\kappa}$ in Lemma 61, when n varies from 2 to 50.

(Note that, in contrast to Lemma 48, there is no $\sqrt{n}$-factor on the right-hand side of the above expression.)

Proof: See Appendix K. $\square$

Let us make a few comments about the choice of $\boldsymbol{\kappa}$ in Lemma 61.

- Figure 8 shows the exact ratios for n from 2 to 50. In particular, note that for $n = 2$ we have
$$\frac{\mathrm{perm}(\mathbf{1}_{2\times 2})}{\mathrm{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\mathbf{1}_{2\times 2})} = 1.$$
- For even integers n and for the choice of $\boldsymbol{\kappa}$ from Lemma 61, the matrix $\boldsymbol{\theta} = \boldsymbol{I}_{(n/2)\times(n/2)} \otimes \mathbf{1}_{2\times 2}$ yields the ratio $\frac{\mathrm{perm}(\boldsymbol{\theta})}{\mathrm{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\theta})} = 1$. This is in stark contrast to Conjecture 53 where $\boldsymbol{\theta}$ represents the conjectured "worst-case" matrix for the ratio $\frac{\mathrm{perm}(\boldsymbol{\theta})}{\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})}$.
- For integers n and k such that k divides n we have
$$(0.922\ldots)^{n/k} \leqslant \frac{\mathrm{perm}(\boldsymbol{\theta})}{\mathrm{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\theta})} \leqslant 1$$
for the matrix $\boldsymbol{\theta} \triangleq \boldsymbol{I}_{(n/k)\times(n/k)} \otimes \mathbf{1}_{k\times k}$.

Let us conclude this section on the fractional Bethe entropy function with a few comments.

- The SPA message update equations in Section V need to be modified so that its fixed points correspond to stationary points of the fractional Bethe free energy, *i.e.*, so that a modified version of the theorem by Yedidia, Freeman, and Weiss [9] holds. In contrast to the SPA message update equations in Section V, the modified SPA message update equations will be such that the right-going messages depend not only on the previous left-going messages but also on the previous right-going messages, and such that the left-going messages depend not only on the previous right-going messages but also on the previous left-going messages. (We omit the details.) Moreover, the convergence analysis in Section V has to be revisited.
- We leave it as an open problem to explore the $\boldsymbol{\kappa}$ parameter space and to find fractional Bethe permanents for which interesting statements can be made, in particular for which a statement like the one in Theorem 49 can be made.

IX. Conjectures

It is an interesting challenge to look at theorems involving permanents and to prove that the theorems still hold if the permanents in these theorems are replaced by Bethe permanents. Let us mention two conjectures along these lines.

A. Perm-Pseudo-Codewords

The following conjecture is based on a theorem in [63] involving permanents of submatrices of a parity-check matrix.

Definition 62 *Let $\mathcal{C}$ be a binary linear code described by a parity-check matrix $\boldsymbol{H} \in \mathbb{F}_2^{m\times n}$, $m < n$. For a size-$(m+1)$ subset $\mathcal{S}$ of the column index set $\mathcal{I}(\boldsymbol{H})$ we define the Bethe perm-vector based on $\mathcal{S}$ to be the vector $\boldsymbol{\omega} \in \mathbb{Z}^n$ with components*
$$\omega_i \triangleq \begin{cases} \mathrm{perm}_{\mathrm{B}}\left(\boldsymbol{H}_{\mathcal{S}\setminus i}\right) & \text{if } i \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases},$$
where $\boldsymbol{H}_{\mathcal{S}\setminus i}$ is the submatrix of $\boldsymbol{H}$ consisting of all the columns of $\boldsymbol{H}$ whose index is in the set $\mathcal{S}\setminus\{i\}$. $\square$

Conjecture 63 *Let $\mathcal{C}$ be a binary linear code described by the parity-check matrix $\boldsymbol{H} \in \mathbb{F}_2^{m\times n}$, $m < n$, let $\mathcal{K}(\boldsymbol{H})$ be the fundamental cone associated with $\boldsymbol{H}$ [49], [50], and let $\mathcal{S}$ be a size-$(m+1)$ subset of $\mathcal{I}(\boldsymbol{H})$. The Bethe perm-vector $\boldsymbol{\omega}$ based on $\mathcal{S}$ is a pseudo-codeword of $\boldsymbol{H}$,* i.e.,
$$\boldsymbol{\omega} \in \mathcal{K}(\boldsymbol{H}), \tag{12}$$
$\square$

B. Permanent-Based Kernels

Based on a result by Cuturi [64], Huang and Jebara [8] made the following conjecture.

Conjecture 64 (Huang and Jebara [8]) *Let n be a positive integer and let $\mathcal{X}$ be a set endowed with a kernel κ. Let $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ and $Y = \{y_1, \ldots, y_n\} \in \mathcal{X}^n$. Then*
$$\kappa_{\mathrm{perm}_{\mathrm{B}}} : \ (X, Y) \mapsto \mathrm{perm}_{\mathrm{B}}\left(\left[\kappa(x_i, y_j)\right]_{1\leqslant i\leqslant n,\, 1\leqslant j\leqslant n}\right)$$
is a positive definite kernel on $\mathcal{X}^n \times \mathcal{X}^n$. $\square$

X. Conclusions

In this paper, we have pursued a graphical-model-based approach to approximating the permanent of a non-negative square matrix, the resulting approximation being called the Bethe permanent. We have seen that the associated functions, like Bethe entropy function and Bethe free energy function, are remarkably well behaved for a graphical model with a non-trivial cycle structure. In that respect, an important part is played by a theorem by Birkhoff and von Neumann (*cf.* Theorem 3). Moreover, the SPA can be used to efficiently find the minimum of the Bethe free energy function and thereby the Bethe permanent. We have also presented a graph-cover-based analysis that gives additional insights into the inner workings of the Bethe permanent, its strengths, and its weaknesses, and we have commented on Bethe-permanent-based upper and lower bounds on the permanent. Along the way we have stated several conjectures and open problems, that, if answered one way or the other, could further elucidate the relationship between the permanent and the Bethe permanent.

Acknowledgments

We gratefully acknowledge Farzad Parvaresh for pointing out to us the papers [23], [24], Krishna Viswanathan for discussions on the permanent of structured matrices, and Adam Yedidia and Misha Chertkov for sharing an early version of their paper [17].

Appendix A
Proof of Theorem 20

Observe that once the concavity of S is established, it is straightforward to verify the claim in the theorem statement that $S(\boldsymbol{\xi}) \geqslant 0$ for all $\boldsymbol{\xi} \in \Pi_{[n]}$. Indeed, because $\Pi_{[n]}$ is a polytope with n vertices, because S takes on the value 0 at each of these vertices, and because S is concave, this statement is true.

Therefore, let us focus on the concavity statement. Clearly, for $n = 2$ the statement can easily be verified and so the rest of this appendix will only discuss the case $n \geqslant 3$.

By definition, a multi-dimensional function is concave if it is a concave function along any straight line in its domain. Towards showing that this is indeed the case for S, let us fix an arbitrary point $\boldsymbol{\xi} \in \Pi_{[n]}$ and an arbitrary direction $\hat{\boldsymbol{\xi}} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that the function $\boldsymbol{\xi}(t) \triangleq \boldsymbol{\xi} + t \cdot \hat{\boldsymbol{\xi}}$ satisfies $\boldsymbol{\xi}(t) \in \Pi_{[n]}$ for a suitable t-interval around 0 (to be defined later). We need to distinguish three different cases that will be discussed separately in the following subsections:

1) The point $\boldsymbol{\xi}$ is in the interior of $\Pi_{[n]}$.
2) The point $\boldsymbol{\xi}$ is at a vertex of $\Pi_{[n]}$.
3) The point $\boldsymbol{\xi}$ is neither in the interior nor at a vertex of $\Pi_{[n]}$.

A. The Point $\boldsymbol{\xi}$ is in the Interior of $\Pi_{[n]}$

It is straightforward to see that the direction vector $\hat{\boldsymbol{\xi}}$ must satisfy

$$\sum_{\ell} \hat{\xi}_{\ell} = 0, \tag{13}$$

otherwise $\boldsymbol{\xi}(t) \in \Pi_{[n]}$ holds only for $t = 0$. Therefore, we assume that (13) is satisfied. Moreover, because $\boldsymbol{\xi} \in \operatorname{interior}(\Pi_{[n]})$, we have $0 < \xi_{\ell} < 1$, $\ell \in [n]$, and we can find an $\varepsilon > 0$ such that $\boldsymbol{\xi}(t) \in \Pi_{[n]}$ for $-\varepsilon \leqslant t \leqslant \varepsilon$. We will now show that the function $t \mapsto S\big(\boldsymbol{\xi}(t)\big)$ is concave at $t = 0$.

We start by computing the first-order derivative

$$\frac{\mathrm{d}}{\mathrm{d}t} S\big(\boldsymbol{\xi}(t)\big) = -\sum_{\ell} \frac{\mathrm{d}}{\mathrm{d}\xi_{\ell}(t)} s\big(\xi_{\ell}(t)\big) \cdot \hat{\xi}_{\ell},$$

and the second-order derivative

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2} S\big(\boldsymbol{\xi}(t)\big) &= \sum_{\ell} \frac{\mathrm{d}^2}{\mathrm{d}\xi_{\ell}(t)^2} s\big(\xi_{\ell}(t)\big) \cdot \hat{\xi}_{\ell}^2 \\ &\overset{\text{(a)}}{=} -\sum_{\ell} \frac{\hat{\xi}_{\ell}^2}{\xi_{\ell}(t)} + \sum_{\ell} \frac{\hat{\xi}_{\ell}^2}{1 - \xi_{\ell}(t)},\end{aligned}$$

where at step (a) we have used Lemma 18. In particular, at $t = 0$ we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} S\big(\boldsymbol{\xi}(t)\big)\bigg|_{t=0} = \sum_{\ell} \delta_{\ell},$$

where δ_{ℓ}, $\ell \in [n]$, is defined as

$$\delta_{\ell} \triangleq -\sum_{\ell} \frac{\hat{\xi}_{\ell}^2}{\xi_{\ell}} + \sum_{\ell} \frac{\hat{\xi}_{\ell}^2}{1 - \xi_{\ell}} = -\sum_{\ell} \hat{\xi}_{\ell}^2 \cdot \frac{1 - 2\xi_{\ell}}{\xi_{\ell}(1 - \xi_{\ell})}. \tag{14}$$

The proof will be finished once we have shown that $\frac{\mathrm{d}^2}{\mathrm{d}t^2} S\big(\boldsymbol{\xi}(t)\big) \leqslant 0$ at $t = 0$, which is equivalent to the condition that

$$\sum_{\ell} \delta_{\ell} \leqslant 0. \tag{15}$$

We show this by separately considering two cases, the first case being $\boldsymbol{\xi} \in \operatorname{interior}(\Pi_{[n]}) \cap [0, 1/2]^n$, the second case being $\boldsymbol{\xi} \in \operatorname{interior}(\Pi_{[n]}) \setminus [0, 1/2]^n$.

The first case, $\boldsymbol{\xi} \in \operatorname{interior}(\Pi_{[n]}) \cap [0, 1/2]^n$, is relatively straightforward. Namely, for all $\ell \in [n]$ we have $0 < \xi_{\ell} \leqslant 1/2$, which implies $1 - 2\xi_{\ell} \geqslant 0$, which in turn implies $\delta_{\ell} \leqslant 0$, and so (15) is satisfied.

The second case, $\boldsymbol{\xi} \in \operatorname{interior}(\Pi_{[n]}) \setminus [0, 1/2]^n$, needs somewhat more work. We start by observing that there is a unique $\ell^* \in [n]$ such that $\xi_{\ell^*} > 1/2$. (Note that there can only be one such $\ell^* \in [n]$ because $\sum_{\ell} \xi_{\ell} = 1$.) Subsequently, $1 - 2\xi_{\ell^*} < 0$ and $1 - 2\xi_{\ell} > 0$, $\ell \neq \ell^*$.

In the following, it is sufficient to consider only directions $\hat{\boldsymbol{\xi}}$ that satisfy $\hat{\xi}_{\ell^*} > 0$ and $\hat{\xi}_{\ell} \leqslant 0$, $\ell \neq \ell^*$, or that satisfy $\hat{\xi}_{\ell^*} < 0$ and $\hat{\xi}_{\ell} \geqslant 0$, $\ell \neq \ell^*$. This follows from contemplating (13) and (14) and from observing that for a given $\boldsymbol{\xi}$ and given directional magnitudes $\big\{|\hat{\xi}_{\ell}|\big\}_{\ell \neq \ell^*}$, the left-hand side of (15) is maximized by a $\hat{\boldsymbol{\xi}}$ that satisfies the conditions that we have just mentioned.[14] From (13) it follows that such direction vectors $\hat{\boldsymbol{\xi}}$ satisfy

$$|\hat{\xi}_{\ell^*}| = \sum_{\ell \neq \ell^*} |\hat{\xi}_{\ell}|. \tag{16}$$

[14] In other words, such a $\hat{\boldsymbol{\xi}}$ produces the "worst-case" left-hand side in (15): if we can show non-positivity for such direction vectors, we have implicitly shown non-positivity for any other direction vector.

Before continuing, let us introduce

$$\delta' \triangleq -\frac{\big(\hat{\xi}_{\ell^*}\big)^2}{\xi_{\ell^*}} + \sum_{\ell \neq \ell^*} \frac{\hat{\xi}_\ell^2}{1-\xi_\ell},$$

$$\delta'' \triangleq +\frac{\big(\hat{\xi}_{\ell^*}\big)^2}{1-\xi_{\ell^*}} - \sum_{\ell \neq \ell^*} \frac{\hat{\xi}_\ell^2}{\xi_\ell}.$$

Note that $\sum_\ell \delta_\ell = \delta' + \delta''$, and so, if we can show that $\delta' \leqslant 0$ and $\delta'' \leqslant 0$ then we have verified the desired result (15).

The fact $\delta' \leqslant 0$ is a consequence of the equation

$$\sum_{\ell \neq \ell^*} \frac{\hat{\xi}_\ell^2}{1-\xi_\ell} \overset{\text{(a)}}{\leqslant} \frac{1}{\xi_{\ell^*}} \sum_{\ell \neq \ell^*} \hat{\xi}_\ell^2 \overset{\text{(b)}}{\leqslant} \frac{1}{\xi_{\ell^*}} \cdot \left(\sum_{\ell \neq \ell^*} |\hat{\xi}_\ell| \right)^2 \overset{\text{(c)}}{=} \frac{\big(\hat{\xi}_{\ell^*}\big)^2}{\xi_{\ell^*}},$$

where step (a) follows from $\boldsymbol{\xi}$ being in $\Pi_{[n]}$, which implies that $\xi_{\ell^*} = 1 - \sum_{\ell' \neq \ell^*} \xi_{\ell'}$, which in turn implies that $\xi_{\ell^*} \leqslant 1 - \xi_\ell$ for all $\ell \neq \ell^*$. Moreover, step (b) follows from a simple inequality and step (c) follows from (16).

The fact $\delta'' \leqslant 0$ is shown as follows. We start by observing that

$$\begin{aligned}
(1-\xi_{\ell^*}) \cdot \left(\sum_{\ell \neq \ell^*} \frac{\hat{\xi}_\ell^2}{\xi_\ell} \right) &\overset{\text{(a)}}{=} \left(\sum_{\ell \neq \ell^*} \xi_\ell \right) \cdot \left(\sum_{\ell \neq \ell^*} \frac{\hat{\xi}_\ell^2}{\xi_\ell} \right) \\
&= \left(\sum_{\ell \neq \ell^*} \sqrt{\xi_\ell}^2 \right) \cdot \left(\sum_{\ell \neq \ell^*} \left(\frac{|\hat{\xi}_\ell|}{\sqrt{\xi_\ell}} \right)^2 \right) \\
&\overset{\text{(b)}}{\geqslant} \left(\sum_{\ell \neq \ell^*} |\hat{\xi}_\ell| \right)^2 \overset{\text{(c)}}{=} (\hat{\xi}_{\ell^*})^2,
\end{aligned}$$

where step (a) follows from $\boldsymbol{\xi}$ being in $\Pi_{[n]}$ (which implies that $\xi_{\ell^*} = 1 - \sum_{\ell \neq \ell^*} \xi_\ell$), where at step (b) we use the Cauchy-Schwarz inequality, and where at step (c) we use (16). Rearranging this inequality, we see that it is equivalent to the inequality $\delta''_\ell \leqslant 0$.

B. *The Point $\boldsymbol{\xi}$ is at a Vertex of $\Pi_{[n]}$*

Clearly, the direction vector $\hat{\boldsymbol{\xi}}$ must satisfy (13). Moreover, because $\boldsymbol{\xi}$ is at a vertex of $\Pi_{[n]}$, there is an $\ell^* \in [n]$ such that $\xi_{\ell^*} = 1$ and $\xi_\ell = 0$, $\ell \neq \ell^*$, and such that $\hat{\xi}_{\ell^*} < 0$ and $\hat{\xi}_\ell \geqslant 0$, $\ell \neq \ell^*$. Then we can find an $\varepsilon > 0$ such that $\boldsymbol{\xi}(t) \in \Pi_{[n]}$ for $0 \leqslant t \leqslant \varepsilon$. We will now show that the function $t \mapsto S\big(\boldsymbol{\xi}(t)\big)$ is concave at $t = 0$.

We start by plugging in the definition of $\boldsymbol{\xi}(t)$ into $S\big(\boldsymbol{\xi}(t)\big)$, *i.e.*,

$$\begin{aligned}
S\big(\boldsymbol{\xi}(t)\big) &= -\sum_\ell \xi_\ell(t) \log\big(\xi_\ell(t)\big) \\
&\quad + \sum_\ell \big(1-\xi_\ell(t)\big) \log\big(1-\xi_\ell(t)\big) \\
&= -(1+t\hat{\xi}_{\ell^*}) \log(1+t\hat{\xi}_{\ell^*}) - \sum_{\ell \neq \ell^*} (t\hat{\xi}_\ell) \log(t\hat{\xi}_\ell) \\
&\quad + (-t\hat{\xi}_{\ell^*}) \log(-t\hat{\xi}_{\ell^*}) + \sum_{\ell \neq \ell^*} (1-t\hat{\xi}_\ell) \log(1-t\hat{\xi}_\ell).
\end{aligned}$$

From this we compute the first-order derivative

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} S\big(\boldsymbol{\xi}(t)\big) &= -\hat{\xi}_{\ell^*} \log(1+t\hat{\xi}_{\ell^*}) - \hat{\xi}_{\ell^*} \\
&\quad - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(t) - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(\hat{\xi}_\ell) - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \\
&\quad - \hat{\xi}_{\ell^*} \log(t) - \hat{\xi}_{\ell^*} \log(-\hat{\xi}_{\ell^*}) - \hat{\xi}_{\ell^*} \\
&\quad - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(1-t\hat{\xi}_\ell) - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \\
&\overset{\text{(a)}}{=} -\hat{\xi}_{\ell^*} \log(1+t\hat{\xi}_{\ell^*}) - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(\hat{\xi}_\ell) \\
&\quad - \hat{\xi}_{\ell^*} \log(-\hat{\xi}_{\ell^*}) - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(1-t\hat{\xi}_\ell), \qquad (17)
\end{aligned}$$

where at step (a) we have used $\sum_\ell \hat{\xi}_\ell = 0$ multiple times. The second-order derivative is then

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} S\big(\boldsymbol{\xi}(t)\big) = -\frac{\hat{\xi}_{\ell^*}^2}{1+t\hat{\xi}_{\ell^*}} + \sum_{\ell \neq \ell^*} \frac{\hat{\xi}_\ell^2}{1-t\hat{\xi}_\ell}.$$

For $t \downarrow 0$ we obtain

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}t^2} S\big(\boldsymbol{\xi}(t)\big)\bigg|_{t\downarrow 0} &= -\hat{\xi}_{\ell^*}^2 + \sum_{\ell \neq \ell^*} \hat{\xi}_\ell^2 \\
&\overset{\text{(a)}}{=} -\left(-\sum_{\ell \neq \ell^*} \hat{\xi}_\ell \right)^2 + \sum_{\ell \neq \ell^*} \hat{\xi}_\ell^2 \\
&\overset{\text{(b)}}{\leqslant} 0,
\end{aligned}$$

where at step (a) we have used (13) and where step (b) follows from a simple inequality and the fact that $\hat{\xi}_\ell \geqslant 0$ for $\ell \neq \ell^*$. Therefore, the function $t \mapsto S\big(\boldsymbol{\xi}(t)\big)$ is concave at $t = 0$.

C. *The Point $\boldsymbol{\xi}$ is Neither in the Interior nor at a Vertex of $\Pi_{[n]}$*

The fact that $\boldsymbol{\xi}$ is neither in the interior nor at a vertex of $\Pi_{[n]}$ means that there is an $\ell^* \in [n]$ such that $0 < \xi_{\ell^*} < 1$. Clearly, the direction vector $\hat{\boldsymbol{\xi}}$ must satisfy (13), plus some additional constraints that are irrelevant for the discussion here. Then we can find an $\varepsilon > 0$ such that $\boldsymbol{\xi}(t) \in \Pi_{[n]}$ for $0 \leqslant t \leqslant \varepsilon$. The concavity of the function $t \mapsto S\big(\boldsymbol{\xi}(t)\big)$ at $t = 0$ follows then from the observation that, for small non-negative t, the second-order derivative of $S\big(\boldsymbol{\xi}(t)\big)$ w.r.t. t is dominated by the second-order derivative of the expression $-\sum_{\ell:\, \xi_\ell = 0,\, \hat{\xi}_\ell > 0} \xi_\ell(t) \log\big(\xi_\ell(t)\big)$, a function that is concave in t.

Appendix B
Proof of Lemma 24

We obtain the expression in the lemma statement by evaluating $S\big(\boldsymbol{\xi}(t)\big)$ and the first-order derivative of $S\big(\boldsymbol{\xi}(t)\big)$ w.r.t. t at $t = 0$. Clearly, $S\big(\boldsymbol{\xi}(t)\big) = 0$ and so we can focus on computing the first-order derivative.

Fortunately, in Appendix A-B we have already computed the first-order derivative for exactly the same setup. Namely,

from (17) we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} S\big(\boldsymbol{\xi}(t)\big) = -\hat{\xi}_{\ell^*} \log(1 + t\hat{\xi}_{\ell^*}) - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(\hat{\xi}_\ell) - \hat{\xi}_{\ell^*} \log(-\hat{\xi}_{\ell^*}) - \sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(1 - t\hat{\xi}_\ell).$$

In the limit $t \downarrow 0$ this simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}t} S\big(\boldsymbol{\xi}(t)\big)\bigg|_{t\downarrow 0} = -\sum_{\ell \neq \ell^*} \hat{\xi}_\ell \log(\hat{\xi}_\ell) + (-\hat{\xi}_{\ell^*}) \log(-\hat{\xi}_{\ell^*}). \quad (18)$$

This can be rewritten as follows

$$\frac{\mathrm{d}}{\mathrm{d}t} S\big(\boldsymbol{\xi}(t)\big)\bigg|_{t\downarrow 0} = |\hat{\xi}_{\ell^*}| \cdot \left(-\sum_{\ell \neq \ell^*} \frac{|\hat{\xi}_\ell|}{|\hat{\xi}_{\ell^*}|} \log\left(\frac{|\hat{\xi}_\ell|}{|\hat{\xi}_{\ell^*}|} \right) \right),$$

where we have used $-\hat{\xi}_{\ell^*} = |\hat{\xi}_{\ell^*}|$, $\hat{\xi}_\ell = |\hat{\xi}_\ell|$, $\ell \neq \ell^*$, and $|\hat{\xi}_{\ell^*}| = \sum_{\ell \neq \ell^*} |\hat{\xi}_\ell|$, *i.e.*, $\sum_{\ell \neq \ell^*} |\hat{\xi}_\ell| / |\hat{\xi}_{\ell^*}| = 1$. This verifies the expressions for $S\big(\boldsymbol{\xi}(t)\big) = 0$ in the lemma statement.

Finally, the non-negativity of the coefficient of t in (4) follows from $|\hat{\xi}_{\ell^*}| \geqslant |\hat{\xi}_\ell|$, $\ell \neq \ell^*$, which is a consequence of the above-mentioned relation $|\hat{\xi}_{\ell^*}| = \sum_{\ell \neq \ell^*} |\hat{\xi}_\ell|$.

Appendix C
Proof of Lemma 25

Clearly we have $\gamma_{i,j} = 1$ if $j = \sigma(i)$ and $\gamma_{i,j} = 0$ otherwise. From the condition that $\hat{\boldsymbol{\gamma}}$ is such that $\boldsymbol{\gamma}(t) \in \Gamma_{n \times n}$ for small non-negative t, it follows that $\sum_j \hat{\gamma}_{i,j} = 0$ for all $i \in \mathcal{I}$ and $\sum_i \hat{\gamma}_{i,j} = 0$ for all $j \in \mathcal{J}$. Moreover, for every $i \in \mathcal{I}$ we have $\hat{\gamma}_{i,j} \leqslant 0$ if $j = \sigma(i)$ and $\hat{\gamma}_{i,j} \geqslant 0$ otherwise. Then

$$\begin{aligned}
&H_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big) \\
&\overset{\text{(a)}}{=} \frac{1}{2} \sum_i S\big(\boldsymbol{\gamma}_i(t)\big) + \frac{1}{2} \sum_j S\big(\boldsymbol{\gamma}_j(t)\big) \\
&\overset{\text{(b)}}{=} -\frac{t}{2} \sum_i \sum_{j \neq \sigma(i)} \hat{\gamma}_{i,j} \log(\hat{\gamma}_{i,j}) + \frac{t}{2} \sum_i (-\hat{\gamma}_{i,\sigma(i)}) \log(-\hat{\gamma}_{i,\sigma(i)}) \\
&\quad - \frac{t}{2} \sum_j \sum_{i \neq \bar{\sigma}(j)} \hat{\gamma}_{i,j} \log(\hat{\gamma}_{i,j}) + \frac{t}{2} \sum_j (-\hat{\gamma}_{\bar{\sigma}(j),j}) \log(-\hat{\gamma}_{\bar{\sigma}(j),j}) \\
&\quad + O(t^2)
\end{aligned}$$

where step (a) follows from Lemma 21 and where at step (b) we have used $S(\boldsymbol{\gamma}_i) = 0$, $S(\boldsymbol{\gamma}_j) = 0$, and (18).

We observe that in the above expression there are exactly two terms for every edge $e = (i,j) \in \mathcal{I} \times \mathcal{J}$. Rewriting these summations such that all the main summations are over $i \in \mathcal{I}$, we obtain

$$\begin{aligned}
&H_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big) \\
&= -t \sum_i \sum_{j \neq \sigma(i)} \hat{\gamma}_{i,j} \log(\hat{\gamma}_{i,j}) + t \sum_i (-\hat{\gamma}_{i,\sigma(i)}) \log(-\hat{\gamma}_{i,\sigma(i)}) \\
&\quad + O(t^2) \\
&\overset{\text{(a)}}{=} t \sum_i |\hat{\gamma}_{i,\sigma(i)}| \cdot \left(-\sum_{j \neq \sigma(i)} \frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{i,\sigma(i)}|} \log\left(\frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{i,\sigma(i)}|} \right) \right) + O(t^2),
\end{aligned}$$

which is the first display equation in the lemma statement. Here, at step (a) we have used $-\hat{\gamma}_{i,\sigma(i)} = |\hat{\gamma}_{i,\sigma(i)}|$, $\hat{\gamma}_{i,j} = |\hat{\gamma}_{i,j}|$, $j \neq \sigma(i)$, and $|\hat{\gamma}_{i,\sigma(i)}| = \sum_{j \neq \sigma(i)} |\hat{\gamma}_{i,j}|$, *i.e.*, $\sum_{j \neq \sigma(i)} |\hat{\gamma}_{i,j}| / |\hat{\gamma}_{i,\sigma(i)}| = 1$.

The non-negativity of the coefficient of t in the above expression follows from $|\hat{\gamma}_{i,\sigma(i)}| \geqslant |\hat{\gamma}_{i,j}|$, $j \neq \sigma(i)$, which is a consequence of the above-mentioned relation $|\hat{\gamma}_{i,\sigma(i)}| = \sum_{j \neq \sigma(i)} |\hat{\gamma}_{i,j}|$.

On the other hand, rewriting these summations such that all the main summations are over $j \in \mathcal{J}$, we obtain the second display equation in the lemma statement.

Appendix D
Proof of Theorem 26

From the assumptions in the theorem statement it follows that $|\hat{\gamma}_{i,\sigma(i)}| = -\hat{\gamma}_{i,\sigma(i)}$ for all $i \in \mathcal{I}$ and that $|\hat{\gamma}_{i,j}| = \hat{\gamma}_{i,j}$ for all $i \in \mathcal{I}$, $j \in \mathcal{J} \setminus \{\sigma(i)\}$ (see also the proof of Lemma 25 in Appendix C). Then,

$$\begin{aligned}
&U_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big) \\
&\overset{\text{(a)}}{=} -\sum_i (1 + t\hat{\gamma}_{i,\sigma(i)}) \log(\theta_{i,\sigma(i)}) - \sum_i \sum_{j \neq \sigma(i)} (t\hat{\gamma}_{i,j}) \log(\theta_{i,j}) \\
&\overset{\text{(b)}}{=} -\sum_i \log(\theta_{i,\sigma(i)}) - t \sum_i \sum_{j \neq \sigma(i)} |\hat{\gamma}_{i,j}| \log\left(\frac{\theta_{i,j}}{\theta_{i,\sigma(i)}} \right) \\
&\overset{\text{(c)}}{=} C - t \sum_i \sum_{j \neq \sigma(i)} |\hat{\gamma}_{i,j}| \log\left(\frac{\theta_{i,j}}{\theta_{i,\sigma(i)}} \right),
\end{aligned}$$

where at step (a) we have used Corollary 15, where at step (b) we have used that $\sum_j \hat{\gamma}_{i,j} = 0$ holds for every $i \in \mathcal{I}$, *i.e.*, that $-\hat{\gamma}_{i,\sigma(i)} = \sum_{j \neq \sigma(i)} \hat{\gamma}_{i,j} = \sum_{j \neq \sigma(i)} |\hat{\gamma}_{i,j}|$ holds for every $i \in \mathcal{I}$, and where at step (c) we have defined $C \triangleq -\sum_i \log(\theta_{i,\sigma(i)})$. (Note that there is no $O(t^2)$ term in the above expressions.) Then

$$\begin{aligned}
&F_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big) \\
&\quad \overset{\text{(a)}}{=} U_{\mathrm{B}}(\boldsymbol{\gamma}) - H_{\mathrm{B}}(\boldsymbol{\gamma}) \\
&\quad \overset{\text{(b)}}{=} C - t \sum_i \sum_{j \neq \sigma(i)} |\hat{\gamma}_{i,j}| \log\left(\frac{\theta_{i,j}}{\theta_{i,\sigma(i)}} \right) \\
&\qquad - t \sum_i |\hat{\gamma}_{i,\sigma(i)}| \cdot \left(-\sum_{j \neq \sigma(i)} \frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{i,\sigma(i)}|} \log\left(\frac{|\hat{\gamma}_{i,j}|}{|\hat{\gamma}_{i,\sigma(i)}|} \right) \right) \\
&\qquad + O(t^2) \\
&\quad \overset{\text{(c)}}{=} C - t \sum_i \sum_{i' \neq i} |\hat{\gamma}_{i,\sigma(i')}| \log\left(\frac{\theta_{i,\sigma(i')}}{\theta_{i,\sigma(i)}} \right) \\
&\qquad - t \sum_i |\hat{\gamma}_{i,\sigma(i)}| \cdot \left(-\sum_{i' \neq i} \frac{|\hat{\gamma}_{i,\sigma(i')}|}{|\hat{\gamma}_{i,\sigma(i)}|} \log\left(\frac{|\hat{\gamma}_{i,\sigma(i')}|}{|\hat{\gamma}_{i,\sigma(i)}|} \right) \right) \\
&\qquad + O(t^2) \\
&\quad \overset{\text{(d)}}{=} C - t \sum_i \sum_{i' \neq i} \underbrace{\mu_i \cdot p_{i,i'}}_{= Q_{i,i'}} \cdot \big[-\log(p_{i,i'}) + T_{i,i'} \big] + O(t^2),
\end{aligned} \quad (19)$$

where at step (a) we have used Corollary 15, where at step (b) we have inserted the above expression for $U_{\mathrm{B}}(\boldsymbol{\gamma})$ and the expression for $H_{\mathrm{B}}(\boldsymbol{\gamma})$ from Lemma 25, where at step (c) we have replaced the summations over $j \in \mathcal{J}$, $j \neq \sigma(i)$, by

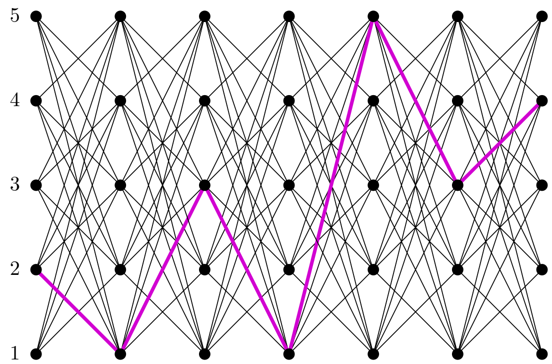


Fig. 9. Trellis for the random walk described in Appendix D. (Here $n = 5$.) Highlighted is an instance of a possible walk.

summations over $i' \in \mathcal{I}$, $\sigma(i') \neq \sigma(i)$, *i.e.*, by summations over $i' \in \mathcal{I}$, $i' \neq i$, and where at step (d) we have introduced the definitions

$$\mu_i \triangleq |\hat{\gamma}_{i,\sigma(i)}| \tag{20}$$

$$p_{i,i'} \triangleq \frac{|\hat{\gamma}_{i,\sigma(i')}|}{|\hat{\gamma}_{i,\sigma(i)}|}, \tag{21}$$

$$Q_{i,i'} \triangleq \mu_i \cdot p_{i,i'} = |\hat{\gamma}_{i,\sigma(i')}|, \tag{22}$$

$$T_{i,i'} \triangleq \log\left(\frac{|\theta_{i,\sigma(i')}|}{|\theta_{i,\sigma(i)}|}\right), \tag{23}$$

for all $(i, i') \in \mathcal{I} \times \mathcal{I}$ with $i \neq i'$. One can verify that the assumptions on $\hat{\boldsymbol{\gamma}}$ imply that

$$\begin{aligned} \sum_i \mu_i &= 1, \\ \sum_{i' \neq i} p_{i,i'} &= 1 \qquad \text{(for all } i \in \mathcal{I}\text{)}, \\ \sum_{i' \neq i} Q_{i,i'} &= \mu_i \qquad \text{(for all } i \in \mathcal{I}\text{)}, \\ \sum_{i \neq i'} Q_{i,i'} &= \mu_{i'} \qquad \text{(for all } i' \in \mathcal{I}\text{)}, \\ \sum_i \sum_{i' \neq i} Q_{i,i'} &= 1. \end{aligned}$$

In order to obtain the theorem statement, we need to maximize the coefficient of $(-t)$ in (19). Before doing this, let us quickly discuss the meaning of this coefficient.

Namely, consider the trellis in Figure 9 with state space $\mathcal{I}$ (*i.e.*, with n states) and where a trellis section has a branch from state $i \in \mathcal{I}$ to state $i' \in \mathcal{I}$ if and only if $i \neq i'$. It is straightforward to see that there is a bijection between, on the one hand, the set of all left-to-right walks in the time-invariant trellis shown in Figure 9, and, on the other hand, the set of backtrackless walks in $\mathsf{N}(\boldsymbol{\theta})$ (*cf.* Figure 1) that were mentioned after Lemma 25. In particular, going from state $i \in \mathcal{I}$ to state $i' \in \mathcal{I} \setminus \{i\}$ in this trellis corresponds to the two half-steps of going from node $i \in \mathcal{I}$ to node $\sigma(i') \in \mathcal{J}$ and then to node $i' \in \mathcal{I}$ in $\mathsf{N}(\boldsymbol{\theta})$. With this, translating (backtrackless) random walks to left-to-right random walks in the trellis in Figure 9, we obtain that

- μ_i is the probability of being in state i,
- $p_{i,i'}$ is the probability of going to state $i' \neq i$, conditioned on being in state i,
- $Q_{i,i'}$ is the probability of being in state i and then going to state $i' \neq i$,
- $-\sum_i \sum_{i' \neq i} \mu_i p_{i,i'} \log(p_{i,i'})$ is the entropy rate of (the Markov chain corresponding to) the random walk on this trellis,
- $T_{i,i'}$ is a branch metric,
- $\sum_i \sum_{i' \neq i} \mu_i p_{i,i'} T_{i,i'}$ is the average branch metric of the random walk on this trellis,
- and maximizing the coefficient of $(-t)$ in the above expression for $F_{\mathrm{B}}\big(\boldsymbol{\gamma}(t)\big)$ means to find the (time-invariant) left-to-right random walk on this trellis that maximizes

  $$\sum_i \sum_{i' \neq i} \mu_i \cdot p_{i,i'} \cdot \big[-\log(p_{i,i'}) + T_{i,i'}\big],$$

  i.e., the sum of the entropy rate and the average branch metric of the random walk. (In statistical physics terms, this expression can be considered to be some negative free energy function.)

The purpose of rewriting the above expression in the way we did, was so that it is very close to the notation used in [65, Lemma 44] that solved exactly the above maximization problem. (Note that related problems were also solved in [66] and [67].)

As was shown in [65, Lemma 44], the maximal value of

$$\sum_i \sum_{i' \neq i} \underbrace{\mu_i \cdot p_{i,i'}}_{=\, Q_{i,i'}} \cdot \big[-\log(p_{i,i'}) + T_{i,i'}\big]$$

is $\log(\rho)$ and is attained by

$$\begin{aligned} \mu_i^* &= \kappa \cdot u_i^{\mathrm{L}} \cdot u_i^{\mathrm{R}}, \\ p_{i,i'}^* &= \begin{cases} \frac{u_{i'}^{\mathrm{R}}}{u_i^{\mathrm{R}}} \cdot \frac{A_{i,i'}}{\rho} & \text{(if } i \neq i'\text{)} \\ 0 & \text{(otherwise)} \end{cases}, \\ Q_{i,i'}^* &= \mu_i^* \cdot p_{i,j}^* = \begin{cases} \kappa \cdot \frac{u_i^{\mathrm{L}} \cdot A_{i,i'} \cdot u_{i'}^{\mathrm{R}}}{\rho} & \text{(if } i \neq i'\text{)} \\ 0 & \text{(otherwise)} \end{cases}, \end{aligned}$$

where $\boldsymbol{A}$, ρ, $\boldsymbol{u}^{\mathrm{L}}$, and $\boldsymbol{u}^{\mathrm{R}}$ are defined in the theorem statement, and where κ is a normalization constant such that $\sum_i \mu_i^* = 1$. Note that $\boldsymbol{A}$, called the noisy adjacency matrix in [65, Lemma 44], is such that $A_{i,i'} = \exp(T_{i,i'})$ for $i \neq i'$ and such that $A_{i,i} = 0$.

Because $\boldsymbol{A}$ contains only non-negative entries, ρ is the so-called Perron eigenvector of $\boldsymbol{A}$, and $\boldsymbol{u}^{\mathrm{L}}$ and $\boldsymbol{u}^{\mathrm{R}}$ are the so-called left and right, respectively, Perron eigenvectors of $\boldsymbol{A}$; one can show that these two vectors contain only non-negative entries.

Translating this result back using (20), (21), and (22), we obtain the result given in the theorem statement.

Appendix E
Proof of Lemma 29

We start by formulating the SPA message update rule for functions node g_i, $i \in \mathcal{I}$, at iteration $t \geqslant 1$. Following [28]–[30], we have for every $i \in \mathcal{I}$, every $j \in \mathcal{J}$, and every $\bar{a}_{i,j} \in$

$\mathcal{A}_{i,j}$,

$$\overrightarrow{\mu}_{i,j}^{(t)}(\bar{a}_{i,j}) \triangleq \frac{1}{C_{i,j}} \cdot \sum_{\substack{\boldsymbol{a}_i \in \mathcal{A}_i \\ a_{i,j}=\bar{a}_{i,j}}} f_i(\boldsymbol{a}_i) \cdot \prod_{j' \neq j} \overleftarrow{\mu}_{i,j'}^{(t-1)}(a_{i,j'}),$$

where $C_{i,j}$ is some suitable normalization constant. Consequently, the update of the likelihood ratio reads

$$\begin{aligned}
\overrightarrow{\Lambda}_{i,j}^{(t)} &\triangleq \frac{\overrightarrow{\mu}_{i,j}^{(t)}(0)}{\overrightarrow{\mu}_{i,j}^{(t)}(1)} = \frac{\sum_{\substack{\boldsymbol{a}_i \in \mathcal{A}_i \\ a_{i,j}=0}} f_i(\boldsymbol{a}_i) \cdot \prod_{j'' \neq j} \overleftarrow{\mu}_{i,j}^{(t-1)}(a_{i,j''})}{\sum_{\substack{\boldsymbol{a}_i \in \mathcal{A}_i \\ a_{i,j}=1}} f_i(\boldsymbol{a}_i) \cdot \prod_{j'' \neq j} \overleftarrow{\mu}_{i,j''}^{(t-1)}(a_{i,j''})} \\
&\overset{\text{(a)}}{=} \frac{\sum_{j' \neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mu}_{i,j'}^{(t-1)}(1) \cdot \prod_{j'' \neq j, j'} \overleftarrow{\mu}_{i,j''}^{(t-1)}(0)}{\sqrt{\theta_{i,j}} \cdot \prod_{j'' \neq j} \overleftarrow{\mu}_{i,j''}^{(t-1)}(0)} \\
&\overset{\text{(b)}}{=} \frac{1}{\sqrt{\theta_{i,j}}} \cdot \sum_{j' \neq j} \sqrt{\theta_{i,j'}} \cdot \left(\overleftarrow{\Lambda}_{i,j'}^{(t-1)} \right)^{-1},
\end{aligned}$$

where at step (a) we have used $\mathcal{A}_i = \{\boldsymbol{u}_j \mid j \in \mathcal{J}\}$ for simplifying the numerator, and where at step (b) we have used the definition of $\overleftarrow{\Lambda}_{i,j'}^{(t-1)}$, $j' \neq j$. This yields the first expression in the lemma statement. The second expression is obtained analogously by considering the SPA message update rule for function nodes g_j, $j \in \mathcal{J}$, at iteration $t \geqslant 1$.

Now we turn our attention to computing the beliefs at the function nodes g_i, $i \in \mathcal{I}$, at iteration $t \geqslant 0$. Following [28]–[30], we have for every $i \in \mathcal{I}$ and every $\boldsymbol{a}_i \in \mathcal{A}_i$,

$$\beta_{i,\boldsymbol{a}_i}^{(t)} = \frac{1}{C_i} \cdot f_i(\boldsymbol{a}_i) \cdot \prod_j \overleftarrow{\mu}_{i,j}^{(t)}(a_{i,j}),$$

where C_i is chosen such that $\sum_{\boldsymbol{a}_i} \beta_{i,\boldsymbol{a}_i}^{(t)} = 1$. In particular, for $\boldsymbol{a}_i = \boldsymbol{u}_j$, $j \in \mathcal{J}$, we get

$$\begin{aligned}
\beta_{i,\boldsymbol{a}_i}^{(t)} &= \frac{1}{C_i} \cdot f_i(\boldsymbol{a}_i) \cdot \prod_{j'} \overleftarrow{\mu}_{i,j'}^{(t)}(a_{i,j'}) \\
&= \frac{1}{C_i} \cdot f_i(\boldsymbol{a}_i) \cdot \left(\prod_{j'} \overleftarrow{\mu}_{i,j'}^{(t)}(0) \right) \cdot \prod_{j'} \frac{\overleftarrow{\mu}_{i,j'}^{(t)}(a_{i,j'})}{\overleftarrow{\mu}_{i,j'}^{(t)}(0)} \\
&= \frac{1}{C_i} \cdot \sqrt{\theta_{i,j}} \cdot \left(\prod_{j'} \overleftarrow{\mu}_{i,j'}^{(t)}(0) \right) \cdot \overleftarrow{V}_{i,j}^{(t)}.
\end{aligned}$$

Because C_i and the expression in the parentheses are independent of j, we have just verified the third expression in the lemma statement. The fourth expression in the lemma statement is obtained analogously by considering the beliefs at function nodes g_j, $j \in \mathcal{J}$, at iteration $t \geqslant 1$.

Appendix F
Proof of Lemma 31

The pseudo-dual function of the Bethe free energy function is given by evaluating the Lagrangian of the Bethe free energy function at a stationary point [37]. Therefore, in a first step, we want to write down the Lagrangian of the Bethe free energy function. To that end, we take the Bethe free energy function as in Definition 10, *i.e.*,

$$\begin{aligned}
F_{\mathrm{B}}(\boldsymbol{\beta}) = &\sum_i U_{\mathrm{B},i}(\boldsymbol{\beta}_i) + \sum_j U_{\mathrm{B},j}(\boldsymbol{\beta}_j) \\
&- \sum_i H_{\mathrm{B},i}(\boldsymbol{\beta}_i) - \sum_j H_{\mathrm{B},j}(\boldsymbol{\beta}_j) + \sum_e H_{\mathrm{B},e}(\boldsymbol{\beta}_e).
\end{aligned}$$

(For the purposes of this appendix, the expression for F_{B} in Definition 10 is somewhat more convenient than the one in Lemma 14.)

Now, introducing a Lagrange multiplier for the edge consistency constraints (but not for the other constraints imposed by the local marginal polytope $\mathcal{B}$, *cf.* Definition 9), we obtain the relevant Lagrangian

$$\begin{aligned}
&L_{\mathrm{Bethe}}\big(\{\boldsymbol{\beta}_i\}, \{\boldsymbol{\beta}_j\}, \{\boldsymbol{\beta}_e\}, \{\overleftarrow{\boldsymbol{\lambda}}_e\}, \{\overrightarrow{\boldsymbol{\lambda}}_e\}\big) \\
&\quad = F_{\mathrm{B}}(\{\boldsymbol{\beta}_i\}, \{\boldsymbol{\beta}_j\}, \{\boldsymbol{\beta}_e\}) \\
&\qquad - \sum_{e=(i,j)} \sum_{a_e} \overleftarrow{\lambda}_{e,a_e} \cdot \left(\sum_{\boldsymbol{a}_i :\, a_{i,e}=a_e} \beta_{i,\boldsymbol{a}_i} - \beta_{e,a_e} \right) \\
&\qquad - \sum_{e=(i,j)} \sum_{a_e} \overrightarrow{\lambda}_{e,a_e} \cdot \left(\sum_{\boldsymbol{a}_j :\, a_{j,e}=a_e} \beta_{j,\boldsymbol{a}_j} - \beta_{e,a_e} \right),
\end{aligned}$$

Because F_{B} is convex in $\{\boldsymbol{\beta}_i\}_i$ and $\{\boldsymbol{\beta}_j\}_j$, but concave in $\{\boldsymbol{\beta}_e\}_e$, the pseudo-dual function of F_{B} is given by

$$\begin{aligned}
&F_{\mathrm{Bethe}}^{\#}\big(\{\overleftarrow{\boldsymbol{\lambda}}_e\}, \{\overrightarrow{\boldsymbol{\lambda}}_e\}, \{\eta_i\}, \{\eta_j\}, \{\eta_e\}\big) \\
&\quad = \max_{\{\boldsymbol{\beta}_e\}} \min_{\{\boldsymbol{\beta}_i\},\, \{\boldsymbol{\beta}_j\}} \\
&\qquad\qquad L_{\mathrm{Bethe}}\big(\{\boldsymbol{\beta}_i\}, \{\boldsymbol{\beta}_j\}, \{\boldsymbol{\beta}_e\}, \\
&\qquad\qquad\qquad \{\overleftarrow{\boldsymbol{\lambda}}_e\}, \{\overrightarrow{\boldsymbol{\lambda}}_e\}, \{\eta_i\}, \{\eta_j\}, \{\eta_e\}\big),
\end{aligned}$$

where the maximization/minimization is over all $\{\boldsymbol{\beta}_e\}_e$, $\{\boldsymbol{\beta}_i\}_i$, $\{\boldsymbol{\beta}_j\}_j$ that satisfy the constraints imposed by the local marginal polytope $\mathcal{B}$, except for the edge consistency constraints. We obtain the maximizing $\{\boldsymbol{\beta}_e\}_e$ and the minimizing $\{\boldsymbol{\beta}_i\}_i$, $\{\boldsymbol{\beta}_j\}_j$ by setting suitable partial derivatives to zero. This yields,

$$\begin{aligned}
\beta_{i,\boldsymbol{a}_i} &= \frac{1}{Z_i} \cdot g_i(\boldsymbol{a}_i) \cdot \prod_{e:\, i(e)=i} \exp\left(\overleftarrow{\lambda}_{e,a_{i,j(e)}} \right), \\
\beta_{j,\boldsymbol{a}_j} &= \frac{1}{Z_j} \cdot g_j(\boldsymbol{a}_j) \cdot \prod_{e:\, j(e)=j} \exp\left(\overrightarrow{\lambda}_{e,a_{i(e),j}} \right), \\
\beta_{e,a_e} &= \frac{1}{Z_e} \cdot \exp\left(\overleftarrow{\lambda}_{e,a_e} \right) \cdot \exp\left(\overrightarrow{\lambda}_{e,a_e} \right),
\end{aligned}$$

where $i(e)$ and $j(e)$ give the label of the, respectively, left and right vertex to which e is incident, and where $\{Z_i\}_i$, $\{Z_j\}_j$, and $\{Z_e\}_e$ are suitable normalization constants such that relevant sums are equal to one.

Now, plugging these beliefs into the Lagrangian, we obtain

(after cancelling several terms) the expression

$$
\begin{aligned}
&F^{\#}_{\mathrm{Bethe}}\big(\{\overleftarrow{\boldsymbol{\lambda}}_e\},\{\overrightarrow{\boldsymbol{\lambda}}_e\}\big)\\
&\quad= -\sum_i \log(Z_i) - \sum_j \log(Z_j) + \sum_e \log(Z_e)\\
&\quad= -\sum_i \log\left(\sum_{\boldsymbol{a}_i} g_i(\boldsymbol{a}_i)\cdot \prod_{e:\, i(e)=i} \exp\left(\overleftarrow{\lambda}_{e,a_{i,j(e)}}\right)\right)\\
&\qquad - \sum_j \log\left(\sum_{\boldsymbol{a}_j} g_j(\boldsymbol{a}_j)\cdot \prod_{e:\, j(e)=j} \exp\left(\overrightarrow{\lambda}_{e,a_{i(e),j}}\right)\right)\\
&\qquad + \sum_e \log\left(\sum_{a_e} \exp\left(\overleftarrow{\lambda}_{e,a_e} + \overrightarrow{\lambda}_{e,a_e}\right)\right).
\end{aligned}
$$

We proceed by using some details of the definition of $\mathsf{N}(\boldsymbol{\theta})$. Namely, using the definition of the local function nodes and taking advantage of the binary alphabet $\mathcal{A}_e = \{0,1\}$, $e \in \mathcal{E}$, we obtain (after some simplifications)

$$
\begin{aligned}
&F^{\#}_{\mathrm{Bethe}}\big(\{\overleftarrow{\boldsymbol{\lambda}}_e\},\{\overrightarrow{\boldsymbol{\lambda}}_e\}\big)\\
&= -\sum_i \log\left(\sum_j \sqrt{\theta_{i,j}}\cdot \exp\left(\overleftarrow{\lambda}_{(i,j),1} - \overleftarrow{\lambda}_{(i,j),0}\right)\right)\\
&\quad - \sum_j \log\left(\sum_i \sqrt{\theta_{i,j}}\cdot \exp\left(\overrightarrow{\lambda}_{(i,j),1} - \overrightarrow{\lambda}_{(i,j),0}\right)\right)\\
&\quad + \sum_e \log\Big(1 + \exp\Big(\big(\overleftarrow{\lambda}_{e,1} - \overleftarrow{\lambda}_{e,0}\big) + \big(\overrightarrow{\lambda}_{e,1} - \overrightarrow{\lambda}_{e,0}\big)\Big)\Big)
\end{aligned}
$$

From the results in [9] it follows that at a fixed point of the SPA, the quantity $\overleftarrow{\lambda}_{(i,j),0} - \overleftarrow{\lambda}_{(i,j),1}$ represents the log-likelihood ratio of the left-going message along the edge (i,j), and the quantity $\overrightarrow{\lambda}_{(i,j),0} - \overrightarrow{\lambda}_{(i,j),1}$ represents the log-likelihood ratio of the right-going message along the edge (i,j). Clearly, for every edge $(i,j) \in \mathcal{I}\times\mathcal{J}$, these quantities are related to the inverse likelihood ratios by

$$
\begin{aligned}
\overleftarrow{\mathrm{V}}_{i,j} &= \exp\left(\overleftarrow{\lambda}_{(i,j),1} - \overleftarrow{\lambda}_{(i,j),0}\right),\\
\overrightarrow{\mathrm{V}}_{i,j} &= \exp\left(\overrightarrow{\lambda}_{(i,j),1} - \overrightarrow{\lambda}_{(i,j),0}\right),
\end{aligned}
$$

respectively. Therefore, we get

$$
\begin{aligned}
&F^{\#}_{\mathrm{Bethe}}\big(\{\overleftarrow{\mathrm{V}}_{i,j}\},\{\overrightarrow{\mathrm{V}}_{i,j}\}\big)\\
&\quad= -\sum_i \log\left(\sqrt{\theta_{i,j}}\cdot \overleftarrow{\mathrm{V}}_{i,j}\right) - \sum_j \log\left(\sqrt{\theta_{i,j}}\cdot \overrightarrow{\mathrm{V}}_{i,j}\right)\\
&\qquad + \sum_{i,j} \log\left(1 + \overleftarrow{\mathrm{V}}_{i,j}\cdot \overrightarrow{\mathrm{V}}_{i,j}\right),
\end{aligned}
$$

which is the expression in the lemma statement.

Although the interpretation of the log-likelihood ratios was given by looking at fixed points of the SPA, it is not difficult to see that we can evaluate this last expression for any set of inverse likelihood ratios.

APPENDIX G
PROOF OF THEOREM 32

This appendix has two subsections. The first subsection considers the case where the global minimum of F_{B} is achieved at a vertex of $\Gamma_{n\times n}$, whereas the second subsection considers the case where the global minimum of F_{B} is achieved in the interior of $\Gamma_{n\times n}$.

For ease of reference, we reproduce here the SPA message update rules from Lemma 29, *i.e.*,

$$
\overrightarrow{\mathrm{V}}^{(t)}_{i,j} = \frac{\sqrt{\theta_{i,j}}}{\sum_{j'\neq j}\sqrt{\theta_{i,j'}}\cdot \overleftarrow{\mathrm{V}}^{(t-1)}_{i,j'}},\ t \geqslant 1,\ (i,j)\in\mathcal{I}\times\mathcal{J}, \tag{24}
$$

$$
\overleftarrow{\mathrm{V}}^{(t)}_{i,j} = \frac{\sqrt{\theta_{i,j}}}{\sum_{i'\neq i}\sqrt{\theta_{i',j}}\cdot \overrightarrow{\mathrm{V}}^{(t)}_{i',j}},\quad t \geqslant 1,\ (i,j)\in\mathcal{I}\times\mathcal{J}. \tag{25}
$$

In both parts of this appendix, the main task will be to exhibit a contraction operation of a suitably chosen subset of the SPA messages.

A. Global Minimum of F_{B} is Achieved at a Vertex of $\Gamma_{n\times n}$

Let $\boldsymbol{\gamma} \in \mathcal{C}_{\mathcal{E}}$ be the vertex of $\Gamma_{n\times n}$ that uniquely minimizes F_{B}. This means that $\boldsymbol{\gamma}$ corresponds to the permutation $\sigma_{\boldsymbol{\gamma}}$. (In the following statement we will use the short-hands $\sigma \triangleq \sigma_{\boldsymbol{\gamma}}$ and $\bar{\sigma} \triangleq \sigma_{\boldsymbol{\gamma}}^{-1}$.)

From (24) it follows that $\overrightarrow{\Lambda}^{(t)}_{i,\sigma(i)} = 1/\overrightarrow{\mathrm{V}}^{(t)}_{i,\sigma(i)}$, $i \in \mathcal{I}$, can be written as[15]

$$
\overrightarrow{\Lambda}^{(t)}_{i,\sigma(i)} = \frac{1}{\sqrt{\theta_{i,\sigma(i)}}}\cdot \sum_{j\neq\sigma(i)} \sqrt{\theta_{i,j}}\cdot \overleftarrow{\mathrm{V}}^{(t-1)}_{i,j},\quad t\geqslant 1,\ i\in\mathcal{I}.
$$

On the other hand, for $i \in \mathcal{I}$ and $j \neq \sigma(i)$ the SPA message update equation in (25) implies

$$
\begin{aligned}
\overleftarrow{\mathrm{V}}^{(t-1)}_{i,j} &= \frac{\sqrt{\theta_{i,j}}}{\sum_{i'\neq i}\sqrt{\theta_{i',j}}\cdot \overrightarrow{\mathrm{V}}^{(t-1)}_{i',j}}\\
&= \frac{\sqrt{\theta_{i,j}}}{\sqrt{\theta_{\bar{\sigma}(j),j}}\cdot \overrightarrow{\mathrm{V}}^{(t-1)}_{\bar{\sigma}(j),j}}\cdot \frac{1}{1+\sum\limits_{i'\neq i,\bar{\sigma}(j)} \frac{\sqrt{\theta_{i',j}}\cdot \overrightarrow{\mathrm{V}}^{(t-1)}_{i',j}}{\sqrt{\theta_{\bar{\sigma}(j),j}}\cdot \overrightarrow{\mathrm{V}}^{(t-1)}_{\bar{\sigma}(j),j}}}\\
&\leqslant \frac{\sqrt{\theta_{i,j}}}{\sqrt{\theta_{\bar{\sigma}(j),j}}\cdot \overrightarrow{\mathrm{V}}^{(t-1)}_{\bar{\sigma}(j),j}}\\
&= \frac{\sqrt{\theta_{i,j}}}{\sqrt{\theta_{\bar{\sigma}(j),j}}}\cdot \overrightarrow{\Lambda}^{(t-1)}_{\bar{\sigma}(j),j},\quad t\geqslant 1,\ i\in\mathcal{I},\ j\neq\sigma(i),
\end{aligned}
$$

where the inequality follows from the fact that all terms in the summation $\sum_{i'\neq i,\bar{\sigma}(j)}$ are non-negative. Then, combining the two above expressions, we obtain

$$
\overrightarrow{\Lambda}^{(t)}_{i,\sigma(i)} \leqslant \sum_{j\neq\sigma(i)} \frac{\theta_{i,j}}{\sqrt{\theta_{i,\sigma(i)}}\sqrt{\theta_{\bar{\sigma}(j),j}}}\cdot \overrightarrow{\Lambda}^{(t-1)}_{\bar{\sigma}(j),j},\quad t\geqslant 1,\ i\in\mathcal{I}.
$$

Rearranging terms, we obtain

$$
\begin{aligned}
\frac{\overrightarrow{\Lambda}^{(t)}_{i,\sigma(i)}}{\sqrt{\theta_{i,\sigma(i)}}} &\leqslant \sum_{j\neq\sigma(i)} \frac{\theta_{i,j}}{\theta_{i,\sigma(i)}}\cdot \frac{\overrightarrow{\Lambda}^{(t-1)}_{\bar{\sigma}(j),j}}{\sqrt{\theta_{\bar{\sigma}(j),j}}}\\
&= \sum_{i'\neq i} \frac{\theta_{i,\sigma(i')}}{\theta_{i,\sigma(i)}}\cdot \frac{\overrightarrow{\Lambda}^{(t-1)}_{i',\sigma(i')}}{\sqrt{\theta_{i',\sigma(i')}}},\quad t\geqslant 1,\ i\in\mathcal{I}.
\end{aligned}
$$

[15] For simplicity, because j does not appear on the left-hand side of this equation, we use j as a summation variable on the right-hand side. This is in contrast to (24) where j appears on the left-hand side and where the summation variable on the right-hand side is j'.

Now, for every $t \geqslant 0$, consider the length-n vector $\vec{\boldsymbol{m}}^{(t)}$ whose i-th entry is $\overrightarrow{\Lambda}^{(t)}_{i,\sigma(i)} / \sqrt{\theta_{i,\sigma(i)}}$. Grouping several of the above inequalities together, we obtain the vector inequality

$$\vec{\boldsymbol{m}}^{(t)} \leqslant \boldsymbol{A} \cdot \vec{\boldsymbol{m}}^{(t-1)}, \quad t \geqslant 1, \tag{26}$$

where the vector inequality has to be understood componentwise, and where the $n \times n$ matrix $\boldsymbol{A}$ was defined in Theorem 26 for the vertex $\boldsymbol{\gamma}$ of $\Gamma_{n\times n}$. Let ρ be the maximal (real) eigenvalue of $\boldsymbol{A}$. Then, Corollary 27 and the assumption that $\boldsymbol{\gamma}$ is the unique minimizer of F_{B} allow us to conclude that $\rho < 1$. However, because $\rho < 1$ implies that all eigenvalues of $\boldsymbol{A}$ have magnitude strictly smaller than 1, the update equation in (26) represents a contraction, and so

$$\left\| \vec{\boldsymbol{m}}^{(t)} \right\|_2 \xrightarrow{t\to\infty} 0.$$

Therefore,

$$\overrightarrow{\Lambda}^{(t)}_{i,\sigma(i)} \xrightarrow{t\to\infty} 0, \quad i \in \mathcal{I}.$$

A similar argument shows that

$$\overleftarrow{\Lambda}^{(t)}_{\bar{\sigma}(j),j} \xrightarrow{t\to\infty} 0, \quad j \in \mathcal{J}.$$

Finally, from (24) and (25) and the above results it follows that

$$\overrightarrow{\mathrm{V}}^{(t)}_{i,j} \xrightarrow{t\to\infty} 0, \quad i \in \mathcal{I},\ j \in \mathcal{J},\ j \neq \sigma(i),$$
$$\overleftarrow{\mathrm{V}}^{(t)}_{i,j} \xrightarrow{t\to\infty} 0, \quad i \in \mathcal{I},\ j \in \mathcal{J},\ j \neq \sigma(i).$$

All these quantities converge to zero exponentially fast.

When F_{B} achieves its minimum in the interior of $\Gamma_{n\times n}$, then we have equality between F_{B} and $F^{\#}_{\mathrm{Bethe}}$ at stationary points of the SPA. However, we also have equality in the present case. Namely, evaluating $F^{\#}_{\mathrm{Bethe}}$ (*cf.* Lemma 31) for the above messages, we obtain

$$F^{\#}_{\mathrm{Bethe}}\big(\{\overleftarrow{\mathrm{V}}^{(t)}_{i,j}\}, \{\overrightarrow{\mathrm{V}}^{(t)}_{i,j}\}\big) \xrightarrow{t\to\infty} -\sum_i \log(\theta_{i,\sigma(i)}),$$

which indeed equals $F_{\mathrm{B}}(\boldsymbol{\gamma})$. From $\rho < 1$ and $F_{\mathrm{B}}(\boldsymbol{\gamma}) = -\log\big(\mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)$ it also follows that

$$\left| \exp\left(-F^{\#}_{\mathrm{Bethe}}\Big(\{\overleftarrow{\mathrm{V}}^{(t)}_{i,j}\}, \{\overrightarrow{\mathrm{V}}^{(t)}_{i,j}\}\Big)\right) - \mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) \right| \leqslant C \cdot \mathrm{e}^{-\nu \cdot t}$$

for suitable constants $C, \nu \in \mathbb{R}_{>0}$.

B. *Global Minimum of F_{B} is Achieved in the Interior of $\Gamma_{n\times n}$*

In Corollary 23 we established that the Bethe free energy function of $\mathsf{N}(\boldsymbol{\theta})$ is convex, *i.e.*, it does not have stationary points besides the global minimum. Therefore, using a theorem by Yedidia, Freeman, Weiss [9], we know that fixed points of the SPA correspond to the global minimum of the Bethe free energy function.

Let $\{\overleftarrow{\mathrm{V}}_{i,j}\}_{i,j}$, $\{\overrightarrow{\mathrm{V}}_{i,j}\}_{i,j}$ be inverse likelihood ratios that constitute a fixed point of the SPA update rules in Lemma 29. As such, these inverse likelihoods must satisfy

$$\overrightarrow{\mathrm{V}}_{i,j} = \frac{\sqrt{\theta_{i,j}}}{\sum_{j'\neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mathrm{V}}_{i,j'}}, \tag{27}$$

$$\overleftarrow{\mathrm{V}}_{i,j} = \frac{\sqrt{\theta_{i,j}}}{\sum_{i'\neq i} \sqrt{\theta_{i',j}} \cdot \overrightarrow{\mathrm{V}}_{i',j}}, \tag{28}$$

for every $(i,j) \in \mathcal{E}$. Note that these SPA fixed point inverse likelihood ratios satisfy $0 < \overrightarrow{\mathrm{V}}_{i,j} < \infty$ and $0 < \overleftarrow{\mathrm{V}}_{i,j} < \infty$, otherwise the assumption that we are dealing with an interior point of $\Gamma_{n\times n}$ would be violated.

It follows from the message gauge invariance mentioned in Remark 30 that, for any positive real number C, the inverse likelihoods $\{C \cdot \overleftarrow{\mathrm{V}}_{i,j}\}_{i,j}$, $\{\frac{1}{C} \cdot \overrightarrow{\mathrm{V}}_{i,j}\}_{i,j}$ also constitute a fixed point of the SPA update rules. We will use this fact later on.

On the other hand, let $\{\overleftarrow{\mathrm{V}}^{(t)}_{i,j}\}_{i,j,t}$, $\{\overrightarrow{\mathrm{V}}^{(t)}_{i,j}\}_{i,j,t}$ be a set of inverse likelihoods obtained by running the SPA on $\mathsf{N}(\boldsymbol{\theta})$ according to the SPA update rules in Lemma 29. In the following, we will not work with $\{\overleftarrow{\mathrm{V}}^{(t)}_{i,j}\}_{i,j,t}$, $\{\overrightarrow{\mathrm{V}}^{(t)}_{i,j}\}_{i,j,t}$ directly, but with $\{\overleftarrow{\varepsilon}^{(t)}_{i,j}\}_{i,j,t}$, $\{\overrightarrow{\varepsilon}^{(t)}_{i,j}\}_{i,j,t}$, which are implicitly defined by the equations

$$\overrightarrow{\mathrm{V}}^{(t)}_{i,j} = \overrightarrow{\mathrm{V}}_{i,j} \cdot \left(1 + \overrightarrow{\varepsilon}^{(t)}_{i,j}\right), \tag{29}$$

$$\overleftarrow{\mathrm{V}}_{i,j} = \overleftarrow{\mathrm{V}}_{i,j} \cdot \left(1 + \overleftarrow{\varepsilon}^{(t)}_{i,j}\right). \tag{30}$$

(Note that $-1 < \overleftarrow{\varepsilon}^{(t)}_{i,j} < \infty$ and $-1 < \overrightarrow{\varepsilon}^{(t)}_{i,j} < \infty$.) Clearly, $\{\overleftarrow{\varepsilon}^{(t)}_{i,j}\}_{i,j,t}$, $\{\overrightarrow{\varepsilon}^{(t)}_{i,j}\}_{i,j,t}$ can be considered to be a "measure" of the distance of the SPA messages to the fixed-point messages. In particular, we have established convergence of the SPA if we can show that these values converge to zero for $t \to \infty$.

In a first step, we express the SPA message update rules in terms of $\{\overleftarrow{\varepsilon}^{(t)}_{i,j}\}_{i,j,t}$ and $\{\overrightarrow{\varepsilon}^{(t)}_{i,j}\}_{i,j,t}$.

Lemma 65 *For the right-going messages it holds that*

$$\overrightarrow{\delta}^{(t)}_{i,j} \triangleq \frac{\sum_{j'\neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mathrm{V}}_{i,j'} \cdot \overleftarrow{\varepsilon}^{(t-1)}_{i,j'}}{\sum_{j'\neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mathrm{V}}_{i,j'}}, \tag{31}$$

$$\overrightarrow{\varepsilon}^{(t)}_{i,j} = -\frac{\overrightarrow{\delta}^{(t)}_{i,j}}{1 + \overrightarrow{\delta}^{(t)}_{i,j}},. \tag{32}$$

For the left-going messages it holds that

$$\overleftarrow{\delta}^{(t)}_{i,j} \triangleq \frac{\sum_{i'\neq j} \sqrt{\theta_{i,j'}} \cdot \overrightarrow{\mathrm{V}}^{(t)}_{i',j} \cdot \overrightarrow{\varepsilon}^{(t)}_{i',j}}{\sum_{i'\neq j} \sqrt{\theta_{i,j'}} \cdot \overrightarrow{\mathrm{V}}^{(t)}_{i',j}}, \tag{33}$$

$$\overleftarrow{\varepsilon}^{(t)}_{i,j} = -\frac{\overleftarrow{\delta}^{(t)}_{i,j}}{1 + \overleftarrow{\delta}^{(t)}_{i,j}}. \tag{34}$$

Proof: Let us establish (32). The expression in (34) then

follows analogously. We compute

$$\begin{aligned}
\overrightarrow{\mathrm{V}}_{i,j} \cdot \left(1 + \overrightarrow{\varepsilon}^{(t)}_{i,j}\right) &\overset{\text{(a)}}{=} \overrightarrow{\mathrm{V}}^{(t)}_{i,j} \\
&\overset{\text{(b)}}{=} \frac{\sqrt{\theta_{i,j}}}{\sum_{j' \neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mathrm{V}}^{(t-1)}_{i,j'}} \\
&\overset{\text{(c)}}{=} \frac{\sqrt{\theta_{i,j}}}{\sum_{j' \neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mathrm{V}}_{i,j'} \cdot \left(1 + \overleftarrow{\varepsilon}^{(t-1)}_{i,j'}\right)} \\
&\overset{\text{(d)}}{=} \frac{\sqrt{\theta_{i,j}}}{\left(\sum_{j' \neq j} \sqrt{\theta_{i,j'}} \cdot \overleftarrow{\mathrm{V}}_{i,j'}\right) \cdot \left(1 + \overrightarrow{\delta}^{(t)}_{i,j}\right)} \\
&\overset{\text{(e)}}{=} \frac{\overrightarrow{\mathrm{V}}_{i,j}}{1 + \overrightarrow{\delta}^{(t)}_{i,j}},
\end{aligned}$$

where at step (a) we have used (29), where at step (b) we have used (24), where at step (c) we have used (30), where at step (d) we have used (31), and where at step (e) we have used (27). Dividing both sides by $\overrightarrow{\mathrm{V}}_{i,j}$, and then subtracting 1 from both sides, yields the expression in (32). □

Note that $\overrightarrow{\delta}^{(t)}_{i,j}$ is a weighted arithmetic average of the error values $\{\overleftarrow{\varepsilon}^{(t-1)}_{i,j'}\}_{j' \neq j}$, and that $\overleftarrow{\delta}^{(t)}_{i,j}$ is a weighted arithmetic average of the error values $\{\overrightarrow{\varepsilon}^{(t)}_{i',j}\}_{i' \neq i}$.

Note also that the expressions in (32) and (34) have the following peculiarity. Namely, solving $\varepsilon = -\delta/(1+\delta)$ for δ we obtain $\delta = -\varepsilon/(1+\varepsilon)$, which is structurally the same expression as the first expression but with the roles of ε and δ interchanged.

Lemma 66 *Fix an iteration number $t \geqslant 1$. Taking advantage of the message gauge invariance that was mentioned in Remark 30, we can rescale the left-going and right-going fixed-point messages such that all $\{\overleftarrow{\varepsilon}^{(t-1)}_{i,j}\}_{i,j}$ are non-negative. With this we define the numbers $\overleftarrow{\varepsilon}^{(t-1)}_{\max} \geqslant 0$ and $\overleftarrow{\varepsilon}^{(t)}_{\max} \geqslant 0$ to be the smallest numbers that satisfy*

$$\begin{aligned}
\overleftarrow{\varepsilon}^{(t-1)}_{i,j} &\leqslant \overleftarrow{\varepsilon}^{(t-1)}_{\max}, \quad (i,j) \in \mathcal{E}, \\
\overleftarrow{\varepsilon}^{(t)}_{i,j} &\leqslant \overleftarrow{\varepsilon}^{(t)}_{\max}, \qquad (i,j) \in \mathcal{E}.
\end{aligned}$$

Then

$$0 \leqslant \overleftarrow{\varepsilon}^{(t)}_{i,j} \leqslant \overleftarrow{\varepsilon}^{(t)}_{\max} \leqslant \overleftarrow{\varepsilon}^{(t-1)}_{\max} \quad (i,j) \in \mathcal{E}.$$

Proof: It follows immediately from (31) that

$$0 \leqslant \overrightarrow{\delta}^{(t)}_{i,j} \leqslant \overleftarrow{\varepsilon}^{(t-1)}_{\max}, \quad (i,j) \in \mathcal{E},$$

and so, because of (32), we have

$$-1 < -\frac{\overleftarrow{\varepsilon}^{(t-1)}_{\max}}{1 + \overleftarrow{\varepsilon}^{(t-1)}_{\max}} \leqslant \overrightarrow{\varepsilon}^{(t)}_{i,j} \leqslant 0, \quad (i,j) \in \mathcal{E}. \tag{35}$$

Using (33), this implies

$$-1 < -\frac{\overleftarrow{\varepsilon}^{(t-1)}_{\max}}{1 + \overleftarrow{\varepsilon}^{(t-1)}_{\max}} < \overleftarrow{\delta}^{(t)}_{i,j} \leqslant 0, \quad (i,j) \in \mathcal{E},$$

and so, because of (34), we have

$$0 \leqslant \overleftarrow{\varepsilon}^{(t)}_{i,j} \leqslant \overleftarrow{\varepsilon}^{(t-1)}_{\max}, \quad (i,j) \in \mathcal{E}. \tag{36}$$

This proves the statement in the lemma. □

This shows that the errors stay bounded but it does not prove convergence yet. (This result is essentially equivalent to the result that is obtained by taking the zero-temperature limit of the contraction coefficient that is computed in the SPA convergence analysis of [18]: the result is a contraction coefficient of 1, which is non-trivial, but not good enough to show that the message update map is a contraction.[16])

It turns out that in order to improve these bounds we have to track the error values over two iteration, *i.e.*, four half iterations. (We suspect that this is related to the fact that the girth of $\mathsf{N}(\boldsymbol{\theta})$, *i.e.*, the length of the shortest cycle of $(\boldsymbol{\theta})$, is 4.)

Lemma 67 *Fix an iteration number $t \geqslant 1$. Taking advantage of the message gauge invariance that was mentioned in Remark 30, we can rescale the left-going and right-going fixed-point messages such that all $\{\overleftarrow{\varepsilon}^{(t-1)}_{i,j}\}_{i,j}$ are non-negative and such that, additionally, $\min_{i,j} \overleftarrow{\varepsilon}^{(t-1)}_{i,j} = 0$. With this, we define the numbers $\overleftarrow{\varepsilon}^{(t-1)}_{\max} \geqslant 0$ and $\overleftarrow{\varepsilon}^{(t+1)}_{\max} \geqslant 0$ to be the smallest numbers that satisfy*

$$\begin{aligned}
\overleftarrow{\varepsilon}^{(t-1)}_{i,j} &\leqslant \overleftarrow{\varepsilon}^{(t-1)}_{\max}, \quad (i,j) \in \mathcal{E}, \\
\overleftarrow{\varepsilon}^{(t+1)}_{i,j} &\leqslant \overleftarrow{\varepsilon}^{(t+1)}_{\max}, \quad (i,j) \in \mathcal{E}.
\end{aligned}$$

Then

$$0 \leqslant \overleftarrow{\varepsilon}^{(t+1)}_{i,j} \leqslant \overleftarrow{\varepsilon}^{(t+1)}_{\max} \leqslant \nu' \cdot \overleftarrow{\varepsilon}^{(t-1)}_{\max} \quad (i,j) \in \mathcal{E},$$

for some constant $0 \leqslant \nu' < 1$ that depends only on $\boldsymbol{\theta}$ and the fixed-point messages $\{\overleftarrow{\mathrm{V}}_{i,j}\}_{i,j}$ and $\{\overrightarrow{\mathrm{V}}_{i,j}\}_{i,j}$, i.e., *$\nu'$ is independent of t.*

Proof: The statement $\overleftarrow{\varepsilon}^{(t+1)}_{i,j} \geqslant 0$, $(i,j) \in \mathcal{E}$ follows from applying Lemma 66 twice. Therefore, we can focus on the proof of $\overleftarrow{\varepsilon}^{(t+1)}_{\max} \leqslant \nu' \cdot \overleftarrow{\varepsilon}^{(t-1)}_{\max}$.

For a given edge $(i,j) \in \mathcal{E}$, we observe that $-\overleftarrow{\varepsilon}^{(t-1)}_{\max}/\big(1 + \overleftarrow{\varepsilon}^{(t-1)}_{\max}\big) \leqslant \overrightarrow{\varepsilon}^{(t)}_{i,j}$ in (35) holds with equality only if $\overleftarrow{\varepsilon}^{(t-1)}_{i,j} = \overleftarrow{\varepsilon}^{(t-1)}_{\max}$ for all edges (i,j') with $j' \neq j$. Similarly, for a given edge $(i,j) \in \mathcal{E}$ we observe that $\overleftarrow{\varepsilon}^{(t)}_{i,j} \leqslant \overleftarrow{\varepsilon}^{(t-1)}_{\max}$ in (36) holds with equality only if $\overrightarrow{\varepsilon}^{(t)}_{i,j} = -\overleftarrow{\varepsilon}^{(t-1)}_{\max}/\big(1 + \overleftarrow{\varepsilon}^{(t-1)}_{\max}\big)$ for all edges (i',j) with $i' \neq i$. This motivates the definition of the following sets where we track the edges for which a strict inequality holds w.r.t. the inequalities just mentioned. Namely, for $t \geqslant 1$ we define

$$\begin{aligned}
\overrightarrow{\mathcal{E}}^{(t)} &\triangleq \left\{ (i,j) \in \mathcal{E} \;\middle|\; \begin{array}{c} \text{there is at least one edge } (i,j'), \\ j' \neq j, \text{ such that } (i,j') \in \overleftarrow{\mathcal{E}}^{(t-1)} \end{array} \right\}, \\
\overleftarrow{\mathcal{E}}^{(t)} &\triangleq \left\{ (i,j) \in \mathcal{E} \;\middle|\; \begin{array}{c} \text{there is at least one edge } (i',j), \\ i' \neq i, \text{ such that } (i',j) \in \overleftarrow{\mathcal{E}}^{(t)} \end{array} \right\}.
\end{aligned}$$

With this, assume that $\overleftarrow{\mathcal{E}}^{(t-1)}$ contains all the edges for which $\overleftarrow{\varepsilon}^{(t-1)}_{i,j} < \overleftarrow{\varepsilon}^{(t-1)}_{\max}$. Clearly, $\overrightarrow{\mathcal{E}}^{(t)}$ then contains all edges (i,j) for which $\overrightarrow{\varepsilon}^{(t)}_{i,j} > -\overleftarrow{\varepsilon}^{(t-1)}_{\max}/\big(1 + \overleftarrow{\varepsilon}^{(t-1)}_{\max}\big)$. Similarly, $\overleftarrow{\mathcal{E}}^{(t)}$ contains all edges (i,j) for which $\overleftarrow{\varepsilon}^{(t)}_{i,j} < \overleftarrow{\varepsilon}^{(t-1)}_{\max}$.

[16]Given the difference in the graphical model in [18] and the graphical model considered here, some care is required when comparing the temperature that is mentioned here and the temperature that is mentioned in Sections II and III.

If $\overleftarrow{\varepsilon}_{\max}^{(t-1)} = 0$ then the lemma is clearly true. So, assume that $\overleftarrow{\varepsilon}_{\max}^{(t-1)} > 0$. Let $\overleftarrow{\mathcal{E}}^{(t-1)}$ contain all edges (i,j) for which $\overrightarrow{\varepsilon}_{i,j}^{(t-1)} < \overleftarrow{\varepsilon}_{\max}^{(t-1)}$. The assumptions in the lemma statement guarantee that there is at least one such edge, namely the edge(s) (i,j) for which $\overrightarrow{\varepsilon}_{i,j}^{(t-1)} = 0$, and so the set $\overleftarrow{\mathcal{E}}^{(t-1)}$ is non-empty. It can then be verified that four half-iterations later we have $\overleftarrow{\mathcal{E}}^{(t+1)} = \mathcal{E}$.

The fact that there is, as mentioned in the lemma statement, a constant ν' that is t-independent and strictly smaller than 1 is then established by the tracking the differences between the left- and the right-hand side in the above-mentioned strict inequalities. This is done with the help of (31) and (33) $\square$

The convergence proof is then completed by applying Lemma 67 repeatedly. One detail needs to be mentioned, though. Namely, if $\min_{i,j} \overleftarrow{\varepsilon}_{i,j}^{(t+1)} > 0$, and a non-trivial re-gauging occurs at the beginning of the next application of Lemma 67, then in this re-gauging process the value of $\max_{i,j} \overleftarrow{\varepsilon}_{i,j}^{(t+1)} > 0$ never increases (in fact, it always decreases).

Finally, we have

$$\left| \exp\!\Big(- F_{\mathrm{Bethe}}^{\#}\Big(\big\{\overleftarrow{\mathrm{V}}_{i,j}^{(t)}\big\}, \big\{\overrightarrow{\mathrm{V}}_{i,j}^{(t)}\big\}\Big)\Big) - \mathrm{perm}_{\mathrm{B}}(\boldsymbol{\theta}) \right| \leqslant C \cdot \mathrm{e}^{-\nu \cdot t}$$

for suitable constants $C, \nu \in \mathbb{R}_{>0}$. This follows from, on the one hand, the fact that when F_{B} achieves its minimum in the interior of $\Gamma_{n\times n}$ then we have equality between F_{B} and $F_{\mathrm{Bethe}}^{\#}$ at stationary points of the SPA [9], and, on the other hand, the above convergence analysis.

Appendix H
Proof of Lemma 48

In a first step we evaluate $\mathrm{perm}(\mathbf{1}_{n\times n})$. Namely, we obtain

$$\mathrm{perm}(\mathbf{1}_{n\times n}) = n! \overset{\text{(a)}}{=} \sqrt{2\pi n} \cdot \left(\frac{n}{\mathrm{e}}\right)^{n} \cdot \big(1 + o(1)\big), \tag{37}$$

where at step (a) we have used Stirling's approximation of $n!$.

In a second step we evaluate $\mathrm{perm}_{\mathrm{B}}(\mathbf{1}_{n\times n})$. From Definitions 11 and 12 it follows that

$$\mathrm{perm}_{\mathrm{B}}(\mathbf{1}_{n\times n}) \triangleq \exp\left(- \min_{\boldsymbol{\gamma}} F_{\mathrm{B}}(\boldsymbol{\gamma}) \right).$$

From Corollary 23 and symmetry considerations it follows that the minimum in the above expression is achieved by $\gamma_{i,j} = 1/n$, $(i,j) \in \mathcal{I} \times \mathcal{J}$. Therefore,

$$\begin{aligned}
&\log\big(\mathrm{perm}_{\mathrm{B}}(\mathbf{1}_{n\times n})\big) \\
&\quad = - F_{\mathrm{B}}(\boldsymbol{\gamma})\big|_{\gamma_{i,j}=1/n,\ (i,j)\in\mathcal{I}\times\mathcal{J}} \\
&\quad \overset{\text{(a)}}{=} - U_{\mathrm{B}}(\boldsymbol{\gamma}) + H_{\mathrm{B}}(\boldsymbol{\gamma})\big|_{\gamma_{i,j}=1/n,\ (i,j)\in\mathcal{I}\times\mathcal{J}} \\
&\quad \overset{\text{(b)}}{=} -n^2 \cdot \frac{1}{n} \cdot \log\left(\frac{1}{n}\right) + n^2 \cdot \left(1 - \frac{1}{n}\right) \cdot \log\left(1 - \frac{1}{n}\right) \\
&\quad = n \cdot \log(n) + n \cdot (n-1) \cdot \log\left(1 - \frac{1}{n}\right) \\
&\quad = n \cdot \log(n) + n \cdot (n-1) \cdot \left(-\frac{1}{n} - \frac{1}{2n^2} + o\left(\frac{1}{n^2}\right)\right) \\
&\quad = n \cdot \log(n) - (n-1) - \frac{n-1}{2n} + o(1) \\
&\quad = n \cdot \log(n) - n + \frac{1}{2} + o(1),
\end{aligned}$$

where at steps (a) and (b) we have used Corollary 15. Consequently,

$$\mathrm{perm}_{\mathrm{B}}(\mathbf{1}_{n\times n}) = \sqrt{\mathrm{e}} \cdot \left(\frac{n}{e}\right)^{n} \cdot \big(1 + o(1)\big). \tag{38}$$

Combining (37) and (38) we obtain the promised result in the lemma statement.

Appendix I
Proof of Conjecture 51 for $\boldsymbol{\theta} = \mathbf{1}_{n\times n}$

Let $\boldsymbol{\theta} = \mathbf{1}_{n\times n}$. In this appendix we prove that for any $M \in \mathbb{Z}_{>0}$ and any $\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M$ it holds that

$$\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right) \leqslant \big(\mathrm{perm}(\boldsymbol{\theta})\big)^{M}. \tag{39}$$

Although the proof is somewhat lengthy, the combinatorial idea behind it is quite straightforward. Moreover, the only inequality that we use is the AM–GM inequality, which says that the arithmetic mean of a list of non-negative real numbers is at least as large as the geometric mean of this list of numbers. Notably, there is no need to use Stirling's approximation of the factorial function.

Towards showing (39), let us fix some positive integer M, fix some collection of permutation matrices $\widetilde{\boldsymbol{P}} = \big\{\widetilde{\boldsymbol{P}}^{(i,j)}\big\}_{i\in\mathcal{I}, j\in\mathcal{J}} \in \widetilde{\Psi}_M$, define $\widetilde{\boldsymbol{\theta}} \triangleq \boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}$ as in Definition 37, and let the row and column index sets of $\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}$ be $\mathcal{I} \times [M]$ and $\mathcal{J} \times [M]$, respectively. With this, it follows from Definition 1 that

$$\mathrm{perm}(\boldsymbol{\theta}) = \sum_{\sigma} \prod_{i\in\mathcal{I}} \theta_{i,\sigma(i)}, \tag{40}$$

$$\mathrm{perm}(\widetilde{\boldsymbol{\theta}}) = \sum_{\widetilde{\sigma}} \prod_{(i,m)\in\mathcal{I}\times[M]} \widetilde{\theta}_{(i,m),\widetilde{\sigma}((i,m))}, \tag{41}$$

where σ ranges over all permutations of the set $\mathcal{I}$ and where $\widetilde{\sigma}$ ranges over all permutations of the set $\mathcal{I} \times [M]$.

Note that, because all entries of $\widetilde{\boldsymbol{\theta}}$ are either equal to zero or to one, the products in (41) evaluate either to zero or to one. Computing $\mathrm{perm}(\widetilde{\boldsymbol{\theta}})$ is therefore equivalent to counting the $\widetilde{\sigma}$'s for which these products evaluate to one. Equivalently, $\mathrm{perm}(\widetilde{\boldsymbol{\theta}})$ equals the number of perfect matchings in the NFG $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$.

Example 68 *Some of the steps of the proof will be illustrated with the help of the NFGs in Figure 3 (which are reproduced in Figure 10 for ease of reference), where $n = 3$ and $M = 4$.*

- *Figure 10(a) shows the NFG $\mathsf{N}(\boldsymbol{\theta})$; $\mathrm{perm}(\boldsymbol{\theta})$ equals the number of perfect matchings in Figure 10(a). Note:* $\mathrm{perm}(\boldsymbol{\theta}) = n!$.
- *If $\widetilde{\boldsymbol{P}} = \big\{\widetilde{\boldsymbol{P}}^{(i,j)}\big\}_{i\in\mathcal{I},j\in\mathcal{J}} = \big\{\widetilde{\boldsymbol{I}}\big\}_{i\in\mathcal{I},j\in\mathcal{J}}$, where $\widetilde{\boldsymbol{I}}$ is the identity matrix of size $M \times M$, then we obtain the M-cover shown in Figure 10(b), which is a "trivial" M-cover of $\mathsf{N}(\boldsymbol{\theta})$;* $\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right)$ *equals the number of perfect matchings in Figure 10(b). Note:* $\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right) = \big(\mathrm{perm}(\boldsymbol{\theta})\big)^{M} = (n!)^{M}$.
- *For a "non-trivial" collection of permutation matrices $\widetilde{\boldsymbol{P}} = \big\{\widetilde{\boldsymbol{P}}^{(i,j)}\big\}_{i\in\mathcal{I},j\in\mathcal{J}}$ we obtain an M-cover like in Figure 10(c);* $\mathrm{perm}\left(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\right)$ *equals the number of perfect matchings in Figure 10(c).* $\square$

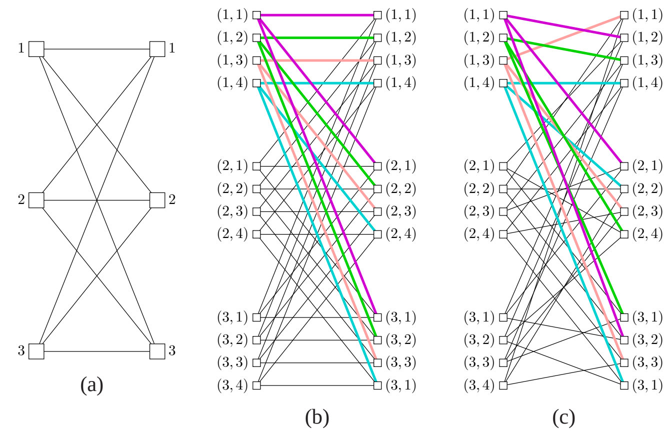


Fig. 10. (a) NFG $\mathsf{N}(\boldsymbol{\theta})$ for $n = 3$. (b) "Trivial" 4-cover of $\mathsf{N}(\boldsymbol{\theta})$ (c) A possible 4-cover of $\mathsf{N}(\boldsymbol{\theta})$. The coloring of the edges in (b) and (c) show visually the fact that he sets $\widetilde{\partial}((i,m))$, $m \in [M]$, form a partition of $\mathcal{J} \times [M]$ (here for $i = 1$). (For more details, see the text in Appendix I).

Let us therefore count the number of perfect matchings in $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$, *cf.* Figure 10(c). Before continuing, we define $\widetilde{\partial}((i,m))$, $(i,m) \in \mathcal{I} \times [M]$, to be the set of neighbors of the vertex (i,m) in $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$, *i.e.*,

$$\widetilde{\partial}((i,m)) \triangleq \left\{ (j,m') \in \mathcal{J} \times [M] \;\middle|\; \widetilde{P}^{(i,j)}_{m,m'} = 1 \right\}.$$

One can easily verify that for every $i \in \mathcal{I}$, the sets $\widetilde{\partial}((i,m))$, $m \in [M]$, form a partition of $\mathcal{J} \times [M]$. (See Figures 10(b)–(c) that highlight this partitioning for $i = 1$.) This observation will be the crucial ingredient of the following steps.

We count the number of perfect matchings in $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$ by considering the vertices $\{(i,m)\}_{m \in [M]}$ for $i = 1$, $i = 2$, up to $i = n$, thereby counting in how many ways we can specify $\widetilde{\sigma}$ such that the product in (41) equals one. Note that because of the above partitioning observation, we can, conditioned on the selection of a perfect matching up to and including step $i-1$ (which we shall symbolically denote by $\widetilde{\sigma}_1^{i-1}$), consider the vertices $\{(i,m)\}_{m \in [M]}$ independently. Then we define

$$\widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}}, \;\; (i,m) \in \mathcal{I} \times [M],$$

to be the number of possibilities of choosing $\widetilde{\sigma}((i,m))$, *i.e.*, the number of ways that the edge of the perfect matching of $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$ that is incident on (i,m) can be chosen.

- Let $i = 1$. Then $\widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}}$, $m \in [M]$, is the number of possibilities of choosing the edge of the perfect matching of $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$ that is incident on (i,m). Because the i-th row of $\boldsymbol{\theta}$ contains only ones, and because of the above partitioning observation, we find that $\widetilde{d}_{i,m} = n$ for all $m \in [M]$, and so,

$$\sum_{m \in [M]} \widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}} = Mn.$$

  We observe that, whatever the selection of these M edges is, M vertices on the right-hand side will be incident on a selected edge, and therefore be "not available anymore" in the following steps. This reduces the number of "available" right-hand side vertices to $Mn - M = M \cdot (n-1)$.
- Let $i = 2$. Then $\widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}}$, $m \in [M]$, is the number of possibilities of choosing the edge of the perfect matching of $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$ that is incident on (i,m). Because the i-th row of $\boldsymbol{\theta}$ contains only ones, because of the above partitioning observation, and because of the observation at the end of the above step, we find that

$$\sum_{m \in [M]} \widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}} \leqslant M \cdot (n-1). \tag{42}$$

  (If all permutation matrices in $\widetilde{\boldsymbol{P}}$ are identity matrices, then it can be verified that the inequality in (42) is an equality. However, for general $\widetilde{\boldsymbol{P}}$, equality in (42) does not need to hold.) Similar to the end of the above step, we observe that whatever the selection of these M edges is, M vertices on the right-hand side will be incident on a selected edge, and therefore be "not available anymore" in the following steps. This reduces the number of "available" right-hand side vertices to $M \cdot (n-1) - M = M \cdot (n-2)$.
- Continuing as above, we observe that for general $i \in \mathcal{I}$ it holds that

$$\sum_{m \in [M]} \widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}} \leqslant M \cdot (n-i+1). \tag{43}$$

Note that for $i \in \mathcal{I}$ we have

$$\begin{aligned} \prod_{m \in [M]} \widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}} &= \left(\prod_{m \in [M]} \widetilde{d}^{\,1/M}_{i,m|\widetilde{\sigma}_1^{i-1}} \right)^M \\ &\overset{\text{(a)}}{\leqslant} \left(\frac{1}{M} \sum_{m \in [M]} \widetilde{d}_{i,m|\widetilde{\sigma}_1^{i-1}} \right)^M \\ &\overset{\text{(b)}}{\leqslant} \left(\frac{1}{M} \cdot M \cdot (n-i+1) \right)^M \\ &= (n-i+1)^M, \end{aligned} \tag{44}$$

where at step (a) we have used the fact that the geometric mean of a collection of non-negative numbers is upper bounded by the arithmetic mean of the same collection of numbers, and where at step (b) we have used (43).

With this, we obtain the following upper bound on $\operatorname{perm}(\widetilde{\boldsymbol{\theta}})$.

Namely,

$$
\begin{aligned}
\operatorname{perm}(\widetilde{\boldsymbol{\theta}}) &\overset{\text{(a)}}{=} \sum_{\widetilde{\sigma}_1^1} \sum_{\widetilde{\sigma}_1^2 | \widetilde{\sigma}_1^1} \cdots \sum_{\widetilde{\sigma}_1^{n-1} | \widetilde{\sigma}_1^{n-2}} \sum_{\widetilde{\sigma}_1^{n} | \widetilde{\sigma}_1^{n-1}} 1 \\
&\overset{\text{(b)}}{=} \sum_{\widetilde{\sigma}_1^1} \sum_{\widetilde{\sigma}_1^2 | \widetilde{\sigma}_1^1} \cdots \sum_{\widetilde{\sigma}_1^{n-1} | \widetilde{\sigma}_1^{n-2}} \prod_{m_n \in [M]} \widetilde{d}_{n, m_n | \widetilde{\sigma}_1^{n-1}} \\
&\overset{\text{(c)}}{\leqslant} \sum_{\widetilde{\sigma}_1^1} \sum_{\widetilde{\sigma}_1^2 | \widetilde{\sigma}_1^1} \cdots \sum_{\widetilde{\sigma}_1^{n-1} | \widetilde{\sigma}_1^{n-2}} (n-n+1)^M \\
&\overset{\text{(d)}}{\leqslant} (n-n+1)^M \cdot \sum_{\widetilde{\sigma}_1^1} \sum_{\widetilde{\sigma}_1^2 | \widetilde{\sigma}_1^1} \cdots \sum_{\widetilde{\sigma}_1^{n-1} | \widetilde{\sigma}_1^{n-2}} 1 \\
&\;\;\vdots \\
&\overset{\text{(e)}}{\leqslant} \prod_{i \in \mathcal{I}} (n-i+1)^M \\
&= (n!)^M \\
&\overset{\text{(f)}}{=} \operatorname{perm}(\boldsymbol{\theta})^M ,
\end{aligned}
$$

where at step (a) we have used the fact that $\operatorname{perm}(\widetilde{\boldsymbol{\theta}})$ equals the number of perfect matchings in $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$, where at step (b) we have used the definition of $\widetilde{d}_{n,m|\widetilde{\sigma}_1^{n-1}}$, where at step (c) we have used (44) for $i = n$, where at step (d) we take advantage of the fact that $(n-n+1)^M$ is independent of $\widetilde{\sigma}_1^{n-1}$, where at step (e) we apply similar results as at steps (b)–(d) (note that for all i, the quantity $(n-i+1)^M$ is independent of $\widetilde{\sigma}_1^{i-1}$), and where at step (f) we have used the observation $\operatorname{perm}(\boldsymbol{\theta}) = n!$. This shows that the desired inequality (39) indeed holds for arbitrary positive integer M and $\widetilde{\boldsymbol{P}} \in \widetilde{\Psi}_M$.

Appendix J
Proof of Lemma 54

We first prove $\operatorname{perm}_{\mathrm{B}}\big(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\big) \geqslant \big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M$ and then $\operatorname{perm}_{\mathrm{B}}\big(\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}\big) \leqslant \big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M$, from which the promised equality follows.

For the rest of the proof, we will use the short-hand $\widetilde{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}^{\uparrow \widetilde{\boldsymbol{P}}}$ and we will assume that there is at least one permutation $\sigma : [n] \to [n]$ such that $\prod_i \theta_{i,\sigma(i)} > 0$ (otherwise, $\operatorname{perm}_{\mathrm{B}}(\widetilde{\boldsymbol{\theta}}) = \operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta}) = 0$). Moreover, $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$ will be the NFG associated with $\widetilde{\boldsymbol{\theta}}$.[17]

Towards proving the first inequality, let $\boldsymbol{\gamma} \in \Gamma_{n \times n}$ be a matrix that minimizes $F_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}$. Based on $\boldsymbol{\gamma}$, we define the $(Mn) \times (Mn)$ matrix $\widetilde{\boldsymbol{\gamma}}$ with entries

$$
\widetilde{\gamma}_{(i,m),(j,m')} \triangleq \gamma_{i,j} \cdot \widetilde{P}^{(i,j)}_{m,m'}
$$

for all $(i,m,j,m') \in \mathcal{I} \times [M] \times \mathcal{J} \times [M]$. One can easily verify that $\widetilde{\boldsymbol{\gamma}} \in \Gamma_{(Mn)\times(Mn)}$ and that $F_{\mathrm{B},\mathsf{N}(\widetilde{\boldsymbol{\theta}})}(\widetilde{\boldsymbol{\gamma}}) = M \cdot F_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}(\boldsymbol{\gamma})$.

[17] Let $\widetilde{\mathsf{N}}$ be the M-cover of $\mathsf{N}(\boldsymbol{\theta})$ corresponding to $\widetilde{\boldsymbol{P}}$. Note that, strictly speaking, $\widetilde{\mathsf{N}}$ and $\mathsf{N}(\widetilde{\boldsymbol{\theta}})$ are not the same NFG. The former is an M-cover of $\mathsf{N}(\boldsymbol{\theta})$ (therefore it has two times Mn function nodes, all of them with degree n), whereas the latter is a complete bipartite graph with two times Mn function nodes. However, with the above condition on $\boldsymbol{\theta}$, for all practical purposes they are the same because $F_{\mathrm{B},\mathsf{N}(\widetilde{\boldsymbol{\theta}})}(\widetilde{\boldsymbol{\gamma}}) < \infty$ only for matrices $\widetilde{\boldsymbol{\gamma}} \in \Gamma_{(Mn)\times(Mn)}$ for which $\widetilde{\gamma}_{(i,m),(j,m')} = 0$ whenever $\widetilde{P}^{(i,j)}_{m,m'} = 0$, $(i,m,j,m') \in \mathcal{I} \times [M] \times \mathcal{J} \times [M]$.

From this and Corollary 15 it then follows that

$$
\operatorname{perm}_{\mathrm{B}}(\widetilde{\boldsymbol{\theta}}) \geqslant \big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M .
$$

Towards proving the second inequality, let $\widetilde{\boldsymbol{\gamma}} \in \Gamma_{(Mn)\times(Mn)}$ be a matrix that minimizes $F_{\mathrm{B},\mathsf{N}(\widetilde{\boldsymbol{\theta}})}$. One can easily verify that $\widetilde{\gamma}_{(i,m),(j,m')} = 0$ whenever $\widetilde{P}^{(i,j)}_{m,m'} = 0$, $(i,m,j,m') \in \mathcal{I} \times [M] \times \mathcal{J} \times [M]$. Based on $\widetilde{\boldsymbol{\gamma}}$, we define the $n \times n$ matrix $\boldsymbol{\gamma}$ with entries

$$
\gamma_{i,j} \triangleq \frac{1}{M} \sum_{m} \sum_{m'} \widetilde{\gamma}_{(i,m),(j,m')} \cdot \widetilde{P}^{(i,j)}_{m,m'}
$$

for all $(i,j) \in \mathcal{I} \times \mathcal{J}$. One can easily verify that $\boldsymbol{\gamma} \in \Gamma_{n \times n}$. Let $\widetilde{\boldsymbol{\gamma}}_{(i,m)}$ be the length-n vector based on the (i,m)-th row of $\widetilde{\boldsymbol{\gamma}}$, where we include an entry only if $\widetilde{P}^{(i,j)}_{m,m'} = 1$. Similarly, define the length-n vector $\widetilde{\boldsymbol{\gamma}}_{(j,m')}$ based on the (j,m')-th column of $\widetilde{\boldsymbol{\gamma}}$. One can verify that the i-th row of $\boldsymbol{\gamma}$, *i.e.*, $\boldsymbol{\gamma}_i$, equals $\frac{1}{M} \sum_m \widetilde{\boldsymbol{\gamma}}_{(i,m)}$. Similarly, the j-th column of $\boldsymbol{\gamma}$, *i.e.*, $\boldsymbol{\gamma}_j$, equals $\frac{1}{M} \sum_{m'} \widetilde{\boldsymbol{\gamma}}_{(j,m')}$. Then

$$
\begin{aligned}
H_{\mathrm{B},\mathsf{N}(\widetilde{\boldsymbol{\theta}})}(\widetilde{\boldsymbol{\gamma}}) &\overset{\text{(a)}}{=} \frac{1}{2} \sum_{i} \sum_{m} S(\widetilde{\boldsymbol{\gamma}}_{(i,m)}) + \frac{1}{2} \sum_{j} \sum_{m'} S(\widetilde{\boldsymbol{\gamma}}_{(j,m')}) \\
&\overset{\text{(b)}}{\leqslant} \frac{M}{2} \sum_{i} S(\widetilde{\boldsymbol{\gamma}}_i) + \frac{M}{2} \sum_{j} S(\widetilde{\boldsymbol{\gamma}}_j) \\
&\overset{\text{(c)}}{=} M \cdot H_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}(\boldsymbol{\gamma}),
\end{aligned}
$$

where at step (a) we have used Lemma 21, where at step (b) we have used the concavity of the S-function (*cf.* Theorem 20), and where at step (c) we have used once again Lemma 21. Moreover, one can easily show that $U_{\mathrm{B},\mathsf{N}(\widetilde{\boldsymbol{\theta}})}(\widetilde{\boldsymbol{\gamma}}) = M \cdot U_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}(\boldsymbol{\gamma})$, and so $F_{\mathrm{B},\mathsf{N}(\widetilde{\boldsymbol{\theta}})}(\widetilde{\boldsymbol{\gamma}}) \geqslant M \cdot F_{\mathrm{B},\mathsf{N}(\boldsymbol{\theta})}(\boldsymbol{\gamma})$. From this and Corollary 15 it then follows that

$$
\operatorname{perm}_{\mathrm{B}}(\widetilde{\boldsymbol{\theta}}) \leqslant \big(\operatorname{perm}_{\mathrm{B}}(\boldsymbol{\theta})\big)^M .
$$

Appendix K
Proof of Lemma 61

Because $\boldsymbol{\kappa}$ satisfies the conditions listed in Theorem 60, the concavity statement for the Bethe entropy function and the convexity statement for the Bethe free energy function follow immediately.

Therefore, let us turn our attention to evaluating the ratio $\operatorname{perm}(\mathbf{1}_{n\times n}) / \operatorname{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\mathbf{1}_{n\times n})$. In a first step we evaluate $\operatorname{perm}(\mathbf{1}_{n\times n})$. Namely, as in Lemma 48, we have

$$
\operatorname{perm}(\mathbf{1}_{n\times n}) = n! = \sqrt{2\pi n} \cdot \left(\frac{n}{\mathrm{e}}\right)^n \cdot \big(1 + o(1)\big). \tag{45}
$$

In a second step we evaluate $\operatorname{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\mathbf{1}_{n\times n})$. From Theorem 60 and symmetry considerations it follows that the minimum in the above expression is achieved by $\gamma_{i,j} = 1/n$,

$(i, j) \in \mathcal{I} \times \mathcal{J}$. Therefore,

$$\begin{aligned}
&\log\left(\mathrm{perm}_{\mathrm{B}}(\mathbf{1}_{n\times n})\right) \\
&\overset{\text{(a)}}{=} -U_{\mathrm{B}}(\boldsymbol{\gamma}) + H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma}) \\
&\overset{\text{(b)}}{=} -n^2 \cdot \left(1 + \frac{1}{2n}\right) \cdot \frac{1}{n} \cdot \log\left(\frac{1}{n}\right) \\
&\qquad + n^2 \cdot \left(1 - \frac{1}{2n}\right) \cdot \left(1 - \frac{1}{n}\right) \cdot \log\left(1 - \frac{1}{n}\right) \\
&= \left(n + \frac{1}{2}\right) \cdot \log(n) + \left(n - \frac{1}{2}\right) \cdot (n-1) \cdot \log\left(1 - \frac{1}{n}\right) \\
&= \left(n + \frac{1}{2}\right) \cdot \log(n) \\
&\qquad + \left(n - \frac{1}{2}\right) \cdot (n-1) \cdot \left(-\frac{1}{n} - \frac{1}{2n^2} + o\left(\frac{1}{n^2}\right)\right) \\
&= \left(n + \frac{1}{2}\right) \cdot \log(n) - n + 1 + o(1),
\end{aligned}$$

where at step (a) we have used $F_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma}) = U_{\mathrm{B}}(\boldsymbol{\gamma}) - H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$, where at (b) we have used $U_{\mathrm{B}}(\boldsymbol{\gamma}) = -\sum_{i,j} \gamma_{i,j} \log(\theta_{i,j}) = 0$ and the expression for $H_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\boldsymbol{\gamma})$ from Lemma 58. Therefore,

$$\mathrm{perm}_{\mathrm{B}}^{(\boldsymbol{\kappa})}(\mathbf{1}_{n\times n}) = \mathrm{e} \cdot \sqrt{n} \cdot \left(\frac{n}{e}\right)^n \cdot \left(1 + o(1)\right). \tag{46}$$

Combining (45) and (46) we obtain the desired result.

REFERENCES

[1] H. Minc, *Permanents*. Reading, MA: Addison-Wesley, 1978.

[2] H. J. Ryser, *Combinatorial Mathematics (Carus Mathematical Monographs No. 14)*. Mathematical Association of America, 1963.

[3] L. Valiant, "The complexity of computing the permanent," *Theor. Comp. Sc.*, vol. 8, no. 2, pp. 189–201, 1979.

[4] P. Dagum and M. Luby, "Approximating the permanent of graphs with large factors," *Theoretical Computer Science*, vol. 102, no. 2, pp. 283–305, 1992.

[5] M. Jerrum, A. Sinclair, and E. Vigoda, "A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries," *J. ACM*, vol. 51, no. 4, pp. 671–697, Jul. 2004.

[6] M. Huber and J. Law, "Fast approximation of the permanent for very dense problems," in *Proc. ACM-SIAM Symp. Discr. Alg.*, San Francisco, CA, USA, Jan. 20–22 2008.

[7] M. Chertkov, L. Kroc, and M. Vergassola, "Belief propagation and beyond for particle tracking," *CoRR, available online under* `http://arxiv.org/abs/0806.1199`, Jun. 2008.

[8] B. Huang and T. Jebara, "Approximating the permanent with belief propagation," *CoRR, available online under* `http://arxiv.org/abs/0908.1769`, Aug. 2009.

[9] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[10] L. Gurvits, "Unharnessing the power of Schrijver's permanental inequality," *CoRR*, `http://arxiv.org/abs/1106.2844`, Jun. 2011.

[11] P. O. Vontobel, "Counting in graph covers: a combinatorial characterization of the Bethe entropy function," *submitted to IEEE Trans. Inf. Theory, available online under* `http://arxiv.org/abs/1012.0065`, Nov. 2010.

[12] R. G. Gallager, *Low-Density Parity-Check Codes*. M.I.T. Press, Cambridge, MA, 1963.

[13] T. Richardson and R. Urbanke, *Modern Coding Theory*. New York, NY: Cambridge University Press, 2008.

[14] Y. Watanabe and M. Chertkov, "Belief propagation and loop calculus for the permanent of a non-negative matrix," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, p. 242002, 2010.

[15] M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, and L. Zdeborová, "Inference in particle tracking experiments by passing messsages between images," *Proc. Natl. Acad. Sci.*, vol. 107, no. 17, pp. 7663–7668, Apr. 2010.

[16] M. Chertkov and V. Y. Chernyak, "Loop series for discrete statistical models on graphs," *J. Stat. Mech.: Theory and Experiment*, p. P06009, Jun. 2006.

[17] A. B. Yedidia and M. Chertkov, "Computing the permanent with belief propagation," *to be submitted to J. Mach. Learning Res., to be posted on arxiv*, Jul. 2011.

[18] M. Bayati and C. Nair, "A rigorous proof of the cavity method for counting matchings," in *Proc. 44th Allerton Conf. on Communications, Control, and Computing*, Allerton House, Monticello, IL, USA, Sep. 27–29 2006.

[19] B. Huang and T. Jebara, "Loopy belief propagation for bipartite maximum weight b-matching," in *Proc. 11th Intern. Conf. on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, Mar. 21–24 2007.

[20] M. Bayati, D. Shah, and M. Sharma, "Max-product for maximum weight matching: convergence, correctness, and LP duality," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1241–1251, Mar. 2008.

[21] S. Sanghavi, D. Malioutov, and A. Willsky, "Belief propagation and LP relaxation for weighted matching in general graphs," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2203–2212, Apr. 2011.

[22] N. Wiberg, "Codes and decoding on general graphs," Ph.D. dissertation, Department of Electrical Engineering, Linköping University, Sweden, 1996.

[23] A. Barvinok, "On the number of matrices and a random matrix with prescribed row and column sums and 0–1 entries," *Adv. in Math.*, vol. 224, no. 1, pp. 316–339, May 2010.

[24] A. Barvinok and A. Samorodnitsky, "Computing the partition function for perfect matchings in a hypergraph," *preprint, available online under* `http://arxiv.org/abs/1009.2397`, Sep. 2010.

[25] J. Yedidia, "An idiosyncratic journey beyond mean field theory," *Advanced Mean Field Methods, Theory and Practice*, pp. 21–36, Jan. 2001.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1990, corrected reprint of the 1985 original.

[28] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[29] G. D. Forney, Jr., "Codes on graphs: normal realizations," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 520–548, Feb. 2001.

[30] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Sig. Proc. Mag.*, vol. 21, no. 1, pp. 28–41, Jan. 2004.

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: John Wiley & Sons Inc., 2006.

[32] P. O. Vontobel, "Connecting the Bethe entropy and the edge zeta function of a cycle code," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, USA, Jun. 13–18 2010, pp. 704–708.

[33] ——, "A factor-graph-based random walk, and its relevance for LP decoding analysis and Bethe entropy characterization," in *Proc. Inf. Theory Appl. Workshop*, UC San Diego, La Jolla, CA, USA, Jan. 31–Feb. 5 2010.

[34] S. Arora, C. Daskalakis, and D. Steurer, "Message-passing algorithms and improved LP decoding," in *Proc. 41st Annual ACM Symp. Theory of Computing*, Bethesda, MD, USA, May 31–June 2 2009.

[35] N. Halabi and G. Even, "LP decoding of regular LDPC codes in memoryless channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 887–897, Feb. 2011.

[36] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[37] P. A. Regalia and J. M. Walsh, "Optimality and duality of the turbo decoder," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1362–1377, Jun. 2007.

[38] M. Mézard and A. Montanari, *Information, Physics, and Computation*. New York, NY: Oxford University Press, 2009.

[39] Y. Weiss and W. T. Freeman, "On the optimality of the max-product belief propagation algorithm in arbitrary graphs," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 736–744, 2001.

[40] ——, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, no. 10, pp. 2173–2200, Oct. 2001.

[41] P. Rusmevichientong and B. Van Roy, "An analysis of belief propagation on the turbo decoding graph with Gaussian densities," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 745–765, 2001.

[42] J. M. Mooij and H. J. Kappen, "Sufficient conditions for convergence of the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4422–4437, Dec. 2007.

[43] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Walk-sums and belief propagation in Gaussian graphical models," *J. Mach. Learn. Res.*, vol. 7, pp. 2031–2064, Dec. 2006.

[44] N. Ruozzi, J. Thaler, and S. Tatikonda, "Graph covers and quadratic minimization," in *Proc. 47th Allerton Conf. on Communications, Control, and Computing*, Allerton House, Monticello, IL, USA, Sep. 30–Oct. 2 2009, pp. 1590–1596.

[45] P. O. Vontobel, "A combinatorial characterization of the Bethe and the Kikuchi partition functions," in *Proc. Inf. Theory Appl. Workshop*, UC San Diego, La Jolla, CA, USA, Feb. 6–11 2011.

[46] W. S. Massey, *Algebraic Topology: an Introduction*. New York: Springer-Verlag, 1977, reprint of the 1967 edition, Graduate Texts in Mathematics, Vol. 56.

[47] H. M. Stark and A. A. Terras, "Zeta functions of finite graphs and coverings," *Adv. in Math.*, vol. 121, no. 1, pp. 124–165, Jul. 1996.

[48] N. L. Biggs, *Discrete Mathematics*, 2nd ed. New York: The Clarendon Press and Oxford University Press, 1989.

[49] R. Koetter and P. O. Vontobel, "Graph covers and iterative decoding of finite-length codes," in *Proc. 3rd Intern. Symp. on Turbo Codes and Related Topics*, Brest, France, Sep. 1–5 2003, pp. 75–82.

[50] P. O. Vontobel and R. Koetter, "Graph-cover decoding and finite-length analysis of message-passing iterative decoding of LDPC codes," *CoRR*, `http://www.arxiv.org/abs/cs.IT/0512078`, Dec. 2005.

[51] A. Schrijver, "Counting 1-factors in regular bipartite graphs," *J. Comb. Theory, Ser. B*, vol. 72, no. 1, pp. 122–135, Jan. 1998.

[52] L. Gurvits, "Van der Waerden / Schrijver-Valiant like conjectures and stable (aka hyperbolic) homogeneous polynomials: one theorem for all," *Elec. J. Comb.*, vol. 15, p. R66, 2008.

[53] M. Laurent and A. Schrijver, "On Leonid Gurvits's proof for permanents," *Amer. Math. Monthly*, vol. 117, no. 10, pp. 903–911, Dec. 2010.

[54] K. Viswanathan, "Pattern maximum-likelihood," Talk at Workshop on "Permanents and modeling probability distributions," American Institute of Mathematics, Palo Alto, CA, USA, Sep. 1 2009.

[55] R. Rivest, "On self-organizing sequential search heuristics," *Comm. ACM*, vol. 19, no. 2, pp. 63–67, Feb. 1976.

[56] J. Sayir, "Ordering memoryless source alphabets using competitive lists," in *Proc. First INTAS International Seminar on Coding Theory and Combinatorics*, Thahkadzor, Armenia, Oct. 6–11 1996.

[57] W. Wiegerinck and T. Heskes, "Fractional belief propagation," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 438–445.

[58] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "A new class of upper bounds on the log partition function," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2313–2335, Jul. 2005.

[59] T. Heskes, "On the uniqueness of loopy belief propagation fixed points," *Neural Computation*, vol. 16, no. 11, pp. 2379–2413, Nov. 2004.

[60] Y. Weiss, T. Meltzer, and C. Yanover, "MAP estimation, linear programming and belief propagation with convex free energies," in *Proc. Conference on Uncertainty in Artificial Intelligence*, Vancouver, Canada, July 19–22 2007.

[61] T. Hazan and A. Shashua, "Norm-product belief propagation: primal-dual message-passing for approximate inference," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6294–6316, Dec. 2010.

[62] N. Ruozzi and S. Tatikonda, "Convergent and correct message passing schemes for optimization problems over graphical models," *submitted to JMLR*, 2010, available online under `http://arxiv.org/abs/1002.3239`.

[63] R. Smarandache and P. O. Vontobel, "Absdet-pseudo-codewords and perm-pseudo-codewords: definitions and properties," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, June 28–July 3 2009.

[64] M. Cuturi, "Permanents, transportation polytopes and positive definite kernels on histograms," in *Proc. Int. Joint Conf. Artificial Intelligence*, Hyderabad, India, Jan. 6–12 2007.

[65] P. O. Vontobel, A. Kavčić, D. M. Arnold, and H.-A. Loeliger, "A generalization of the Blahut-Arimoto algorithm to finite-state channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1887–1918, May 2008.

[66] J. Justesen and T. Høholdt, "Maxentropic Markov chains," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 665–667, Jul. 1984.

[67] A. S. Khayrallah and D. L. Neuhoff, "Coding for channels with cost constraints," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 854–867, May 1996.